

Learning and Prediction Theory of Distributed Least Squares

Siyu Xie, Yaqi Zhang, and Lei Guo, *Fellow, IEEE*

Abstract—With the fast development of the sensor and network technology, distributed estimation has attracted more and more attention, due to its capability in securing communication, in sustaining scalability, and in enhancing safety and privacy. In this paper, we consider a least-squares (LS)-based distributed algorithm build on a sensor network to estimate an unknown parameter vector of a dynamical system, where each sensor in the network has partial information only but is allowed to communicate with its neighbors. Our main task is to generalize the well-known theoretical results on the traditional LS to the current distributed case by establishing both the upper bound of the accumulated regrets of the adaptive predictor and the convergence of the distributed LS estimator, with the following key features compared with the existing literature on distributed estimation: Firstly, our theory does not need the previously imposed independence, stationarity or Gaussian property on the system signals, and hence is applicable to stochastic systems with feedback control. Secondly, the cooperative excitation condition introduced and used in this paper for the convergence of the distributed LS estimate is the weakest possible one, which shows that even if any individual sensor cannot estimate the unknown parameter by the traditional LS, the whole network can still fulfill the estimation task by the distributed LS. Moreover, our theoretical analysis is also different from the existing ones for distributed LS, because it is an integration of several powerful techniques including stochastic Lyapunov functions, martingale convergence theorems, and some inequalities on convex combination of nonnegative definite matrices.

Index Terms—Least squares, distributed estimation, learning, prediction, diffusion strategies, cooperative excitation, regret, martingale theory

I. INTRODUCTION

Distributed estimation algorithms are usually built on a given sensor network for a complex system, aiming at estimating an unknown global system parameter vector cooperatively by the distributed sensors. Each sensor in the network is taken as a node which can only observe partial data of the whole system, perform processing individually, and communicate information only with its neighbors, where the neighbors are defined by the network topology. In recent years, distributed estimation over sensor networks has received increasing research attention, and has been widely studied and used in many areas, e.g., collaborative spectral sensing in cognitive radio

systems, target localization in biological networks, environmental monitoring, military surveillance, and so on (see e.g. [1], [2]). Unlike the traditional centralized method, no node in the network needs to transfer its information to a fusion center for processing in the distributed case, which is more robust and scalable since the fusion center in the centralized method is sensitive and vulnerable to outside attacks. Once the fusion center is under attack, the entire network could collapse. In the distributed method, each node in the network can only exchange data with its neighbors, which may make the communication over the network possible, enhance the safety and privacy of the system, improve the estimation performance, and increase the robustness and scalability of the system (see e.g. [2], [3]).

It goes without saying that different cooperation strategies will lead to different distributed estimation algorithms. For example, the proposed incremental [3]–[6], consensus [7]–[16], and diffusion [17]–[29] strategies, may be combined with different estimation algorithms, e.g., least mean squares (LMS), LS and Kalman filters (KF) [30]–[33], to give rise to different distributed estimation algorithms. Stability and performance analyses have also been established for different distributed estimation algorithms, for example, incremental LMS [3], [4], consensus LMS [7], [8], diffusion LMS [17]–[22], incremental LS [5], [6], consensus LS [9]–[11], diffusion LS [23]–[29], and distributed KF [12]–[16]. In our recent work (see e.g. [7], [8], [17]), we have given the stability and performance results for the consensus and diffusion LMS filters, without imposing the usual independence and stationarity assumptions for the system signals.

Note that the LS is a most basic, widely used and comprehensively studied estimation algorithm in many fields of science and engineering. Moreover, when the unknown parameter is time-invariant, the LS algorithm may generate more accurate estimates in the transient phase and have faster convergence speed compared with LMS algorithm. So the LS appears to be more suitable for applications that require fast speed and accurate estimates for unknown constant parameters. This is one of the main motivations for us to consider the LS-based distributed estimation algorithm in this paper. Another reason for us to study this problem is that the existing convergence theory in the literature is far from satisfactory since it can hardly be applied to non-independent and non-stationary signals coming from practical complex systems where feedback loops inevitably exist.

In fact, almost all the existing studies on the distributed LS (see e.g., [5], [6], [9]–[11], [23]–[29]) require some independent, stationary, or Gaussian assumptions for the system

S. Y. Xie is with the Department of Electrical and Computer Engineering, Wayne State University, Detroit, MI 48202, USA. Email: syxie@wayne.edu.

Y. Q. Zhang and L. Guo are with Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, P. R. China. They are also with School of Mathematical Science, University of Chinese Academy of Sciences, Beijing 100049, P. R. China. Email : zhangyq@amss.ac.cn., lguo@amss.ac.cn.

This work was supported by the National Natural Science Foundation of China under grants 11688101 and 61227902.

signals. For examples, an incremental LS estimation strategy was proposed in [5], and the mean-square performance was studied for independent regressors. Moreover, [9] presented a distributed LS algorithm, and gave stability and performance analyses for independent noises and regressors. [23] proposed a diffusion scheme for LS estimation problem, and analyzed its mean-square convergence under independence conditions on both the system signals and Gaussian noises. Furthermore, [24] presented a diffusion LS algorithm, and proved that the algorithm is asymptotically unbiased and stable for independent regressors and Gaussian noises. In [25], a diffusion bias-compensated LS algorithm was developed, and the closed-form expressions for the residual bias and the mean-square deviation of the estimates were provided under independence and stationarity assumptions. In addition, partial diffusion LS algorithms were proposed in [26], [27], and the performance results were established for ergodic signals [26] and independent signals [27]. Moreover, [28] proposed a reduced communication diffusion LS algorithm for distributed estimation over multi-agent, and [29] developed robust diffusion LS algorithms to mitigate the performance degradation in the presence of impulsive noise. They both established the performance results under independent signal assumptions. Some other related papers, e.g., [6], [10], [11], verified the efficiency of the LS-type algorithms via numerical simulations. All of these indicate that to substantially relax the widely imposed independence and stationarity conditions on the system signals in the analyses of distributed LS, will inevitably bring challenging difficulties in establishing a convergence theory.

Fortunately, in the traditional single sensor case, there is a vast literature on the convergence theory of the classical LS, which is indeed applicable to stochastic systems with feedback control. In fact, motivated by the need to establish a rigorous theory for the well-known LS-based self-tuning regulators proposed by Åström and Wittenmark [34] in stochastic adaptive control, the convergence study of LS with possible stochastic feedback signals had received a great deal of attention in the literature, see e.g., [33], [35]–[42]. At the same time, much effort had also been devoted to stochastic adaptive control, see e.g. [39], [41], [43]–[45]. Among the many significant contributions in this direction, here we only mention that Lai and Wei [38] established a celebrated convergence result under a weakest possible decaying excitation condition on the system signals, and Guo and Chen [42] and Guo [33] finally resolved the longstanding problem concerning the global stability and convergence of the LS-based self-tuning regulators. We remark that the analysis methods including stochastic Lyapunov functions and martingale convergence theorems, which are so useful for the analysis of the classical LS, will also be instrumental for us in investigating the distributed LS algorithm in the current paper.

In this paper, we will provide a theoretical analysis for a distributed LS algorithm of diffusion type [13]–[15], where the diffusion strategy is designed via the so called covariance intersection fusion rule (see, e.g., [46], [47]). In such a diffusion strategy, each node is only allowed to communicate with its neighbors, and both the estimates of the unknown parameter

and the inverse of the covariance matrices are diffused between neighboring nodes. We will generalize the well-known convergence results on the classical LS by establishing both the upper bound of the accumulated regrets of the adaptive predictor and the convergence of the distributed LS estimator, with the following key features compared with the related results in the existing literature:

- Our theory does not need the usually assumed independence, stationarity or Gaussian property on the system signals, and hence does not exclude the applications of the theory to stochastic feedback systems, and will also make it possible for further investigation on related problems concerning the combination of learning, communication and control.
- Our theory for the convergence of the distributed LS is established under a weakest possible cooperative excitation condition which is a natural extension of the single sensor case. The cooperative excitation condition introduced in this paper implies that even if any individual sensor is not able to estimate the unknown parameter, the distributed LS can still accomplish the estimation task. It is also considerably weaker than the related cooperative information condition introduced in our previous work for the theory of the distributed LMS filters (see e.g. [7], [8], [17]).
- The mathematical techniques used in our theoretical analysis are also different from the existing ones for distributed LS. Besides using the powerful techniques from the analysis of the classical LS, we also need to establish some inequalities on convex combination of nonnegative definite matrices and to use the Ky Fan convex theorem [48].

The rest of the paper is organized as follows. In Section II, we present some preliminaries on notations and graph theory, the observation model, and the distributed LS algorithm studied in the paper. The main results are stated in Section III. In Section IV, we provide the proofs of the main results. Finally, some concluding remarks are given in Section V.

II. PROBLEM FORMULATION

A. Basic Notations

In the sequel, $X \in \mathbb{R}^n$ is viewed as an n -dimensional column vector and $A \in \mathbb{R}^{m \times n}$ is viewed as an $m \times n$ -dimensional matrix. Let $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times n}$ be two symmetric matrices, then $A \geq B$ ($A > B$) means $A - B$ is a positive semidefinite (definite) matrix. Also, let $\lambda_{\max}\{\cdot\}$ and $\lambda_{\min}\{\cdot\}$ denote the largest and the smallest eigenvalues of the corresponding matrix respectively. For any matrix $X \in \mathbb{R}^{m \times n}$, $\|X\|$ denotes the operator norm induced by the Euclidean norm, i.e., $(\lambda_{\max}\{XX^T\})^{\frac{1}{2}}$, where $(\cdot)^T$ denotes the transpose operator. We use $\mathbb{E}[\cdot]$ to denote the mathematical expectation operator, and $\mathbb{E}[\cdot|\mathcal{F}_k]$ to denote the conditional mathematical expectation operator, where $\{\mathcal{F}_k\}$ is a sequence of nondecreasing σ -algebras [49]. We also use $\log(\cdot)$ to denote the natural logarithm function, and $\text{Tr}(\cdot)$ to denote the trace of the corresponding matrix. Through out the paper, $|\cdot|$ denotes the determinant of the corresponding matrix, which should

not be confused with the absolute value of a scalar from the context.

Let $\{A_k, k \geq 0\}$ be a matrix sequence and $\{b_k, k \geq 0\}$ be a positive scalar sequence. Then by $A_k = O(b_k)$ we mean that there exists a constant $M > 0$ such that $\|A_k\| \leq Mb_k, \forall k \geq 0$, and by $A_k = o(b_k)$ we mean that $\lim_{k \rightarrow \infty} \|A_k\|/b_k = 0$.

B. Graph Theory

As usual, let the communication structure among sensors be represented by an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, where $\mathcal{V} = \{1, 2, \dots, n\}$ is the set of sensors and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of edges. The structure of the graph \mathcal{G} is described by $\mathcal{A} = \{a_{ij}\}_{n \times n}$ which is called the weighted adjacency matrix, where $a_{ij} > 0$ if $(i, j) \in \mathcal{E}$ and $a_{ij} = 0$ otherwise. Note that $(i, j) \in \mathcal{E} \Leftrightarrow a_{ij} > 0$. In this paper, we assume that the elements of the weighted matrix \mathcal{A} satisfy $a_{ij} = a_{ji}, \forall i, j = 1, \dots, n$, and $\sum_{j=1}^n a_{ij} = 1, \forall i = 1, \dots, n$. Thus the matrix \mathcal{A} is symmetric and doubly stochastic¹.

A path of length ℓ in the graph \mathcal{G} is a sequence of nodes $\{i_1, \dots, i_\ell\}$ subject to $(i_j, i_{j+1}) \in \mathcal{E}$, for $1 \leq j \leq \ell - 1$. The maximum value of the distances between any two nodes in the graph \mathcal{G} is called the diameter of \mathcal{G} . Here in this paper, we assume that the graph is connected, and denote the diameter of the graph \mathcal{G} as $D_{\mathcal{G}}$. Then $1 \leq D_{\mathcal{G}} < \infty$ holds. The set of neighbors of the sensor i is denoted as

$$\mathcal{N}_i = \{j \in \mathcal{V} | (j, i) \in \mathcal{E}\},$$

and the sensor i can only share information with its neighboring sensors from \mathcal{N}_i .

C. Observation Model

Let us consider a sensor network consisting of n sensors. Assume that at each time instant k , each sensor $i \in \{1, \dots, n\}$ in the sensor network receives a noisy scalar measurement $y_{k+1,i}$ and an m -dimensional regressor $\varphi_{k,i} \in \mathbb{R}^m$. They are related by a typical linear stochastic regression model

$$y_{k+1,i} = \varphi_{k,i}^T \boldsymbol{\theta} + w_{k+1,i}, \quad k \geq 0, \quad (1)$$

where $w_{k+1,i}$ is a random noise process, and $\boldsymbol{\theta} \in \mathbb{R}^m$ is an unknown parameter vector which needs to be estimated. Here we assume that at any sensor $i \in \{1, \dots, n\}$, $\varphi_{k,i}$ is \mathcal{F}_k -measurable, where $\{\mathcal{F}_k\}$ is a sequence of nondecreasing σ -algebras. Many problems from different application areas can be cast as (1), see, e.g., [3], [30], [31]. At any time instant $k \geq 1$, sensor i uses both the observations $y_{j+1,i}$ and the regressors $\varphi_{j,i} (j \leq k)$ to estimate the unknown parameter $\boldsymbol{\theta}$, which can be regarded as a supervised learning problem [50].

Because of its ‘‘optimality’’ and fast convergence rate, the well-known LS algorithm is one of the most basic and widely used algorithms in science and technology. The LS estimate at each sensor i is defined by the following at each time instant k :

$$\boldsymbol{\theta}_{k,i} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^m} \sum_{j=1}^k (y_{j,i} - \varphi_{j-1,i}^T \boldsymbol{\theta})^2,$$

¹A matrix is called doubly stochastic, if all elements are nonnegative, both the sum of each row and the sum of each column equal to 1.

which can be solved explicitly and can be calculated recursively as follows (see e.g. [41]):

$$\boldsymbol{\theta}_{k+1,i} = \boldsymbol{\theta}_{k,i} + b_{k,i} P_{k,i} \varphi_{k,i} (y_{k+1,i} - \varphi_{k,i}^T \boldsymbol{\theta}_{k,i}), \quad (2)$$

$$P_{k+1,i} = P_{k,i} - b_{k,i} P_{k,i} \varphi_{k,i} \varphi_{k,i}^T P_{k,i}, \quad (3)$$

$$b_{k,i} = (1 + \varphi_{k,i}^T P_{k,i} \varphi_{k,i})^{-1}, \quad (4)$$

where the initial estimate $\boldsymbol{\theta}_{0,i} \in \mathbb{R}^m$, and the initial positive definite matrix $P_{0,i} \in \mathbb{R}^{m \times m}$ can be chosen arbitrarily. Note that in practice $P_{0,i}$ is usually set as $\alpha_0 I_m$, where α_0 is a positive constant, and I_m denotes the $m \times m$ -dimensional identity matrix.

The above defined LS algorithm can be used for adaptive prediction problems. For any $i \in \{1, \dots, n\}$, and at any time instant $k \geq 1$, the best prediction to the future observation $y_{k+1,i}$ is the following conditional mathematical expectation:

$$\mathbb{E}[y_{k+1,i} | \mathcal{F}_k] = \varphi_{k,i}^T \boldsymbol{\theta},$$

if the noise is a martingale difference sequence with second moment. Unfortunately, this optimal predictor is unavailable because $\boldsymbol{\theta}$ is unknown. A natural way is to construct an adaptive predictor $\hat{y}_{k+1,i}$ by using the online LS estimate $\boldsymbol{\theta}_{k,i}$, i.e.,

$$\hat{y}_{k+1,i} = \varphi_{k,i}^T \boldsymbol{\theta}_{k,i}.$$

The error between the best predictor and the adaptive predictor may be referred to as the regret denoted by

$$R_{k,i} = (\mathbb{E}[y_{k+1,i} | \mathcal{F}_k] - \hat{y}_{k+1,i})^2, \quad (5)$$

which may not be zero and even may not be small in sample paths due to the persistent disturbance of the unpredictable noises in the model. However, one may evaluate the averaged regrets defined as follows:

$$\frac{1}{nt} \sum_{i=1}^n \sum_{k=0}^t R_{k,i}, \quad (6)$$

which we are going to show tends to zero as t increases to infinity under essentially no excitation conditions on the regressors and no independence, stationarity or Gaussian assumptions on system signals, see *Theorem 3.2* below. This is a celebrated property that is widely studied in distributed online learning and optimization problems [51]–[54], but under rather restrictive assumptions such as boundedness, stationarity or independence on the system signals. Moreover, different from [51]–[54], to make the supervised learning result applicable to prediction or classification problem with unseen data, one needs the so called generalization ability in theory, which in turn needs to further study the convergence of the LS estimate itself.

It is well-known that the estimation error of the above classical LS has the following upper bound (see [33], [38]) for each sensor $i \in \{1, \dots, n\}$ as $k \rightarrow \infty$:

$$\|\boldsymbol{\theta}_{k+1,i} - \boldsymbol{\theta}\|^2 = O\left(\frac{\log\left(\lambda_{\max}\{P_{0,i}^{-1}\} + \sum_{j=0}^k \|\varphi_{j,i}\|^2\right)}{\lambda_{\min}\{P_{0,i}^{-1} + \sum_{j=0}^k \varphi_{j,i} \varphi_{j,i}^T\}}\right), \text{ a.s.} \quad (7)$$

Consequently, it is easy to see that the LS estimates will converge to the true parameter if

$$\lim_{k \rightarrow \infty} \frac{\log \left(\lambda_{\max} \{ P_{0,i}^{-1} \} + \sum_{j=0}^k \|\varphi_{j,i}\|^2 \right)}{\lambda_{\min} \left\{ P_{0,i}^{-1} + \sum_{j=0}^k \varphi_{j,i} \varphi_{j,i}^T \right\}} = 0, \quad a.s. \quad (8)$$

Moreover, examples can be constructed to show that if the above limit is a nonzero constant, then the LS estimate cannot converge to the true parameter (see [38]). In this sense, one can say that the condition (8) is the weakest possible one for convergence of the classical LS [38]. Despite of this, the verification of (8) is still a very challenging issue for stochastic adaptive control systems (see e.g. [33], [34], [38], [39], [41]). Moreover, for high-dimensional or sparse stochastic regressors, the condition (8) may indeed be not satisfied. This situation may be improved by exchanging information among nodes in a sensor network on which the distributed LS is defined.

D. Distributed LS Algorithm

In this paper, we will consider the following basic class of distributed LS algorithms of diffusion type, and our main contribution is to establish a convergence theory for general correlated, non-stationary and non-Gaussian regression signals, so that the theory is applicable to control systems.

Algorithm 1 Distributed LS algorithm

For any given sensor $i \in \{1, \dots, n\}$, begin with an initial estimate $\theta_{0,i} \in \mathbb{R}^m$, and an initial positive definite matrix $P_{0,i} \in \mathbb{R}^{m \times m}$. The algorithm is recursively defined at any iteration $k \geq 0$ as follows:

- 1: Adapt (generate $\bar{\theta}_{k+1,i}$ and $\bar{P}_{k+1,i}$ on the bases of $\theta_{k,i}$, $P_{k,i}$ and $\varphi_{k,i}$, $y_{k+1,i}$):

$$\bar{\theta}_{k+1,i} = \theta_{k,i} + b_{k,i} P_{k,i} \varphi_{k,i} (y_{k+1,i} - \varphi_{k,i}^T \theta_{k,i}), \quad (9)$$

$$\bar{P}_{k+1,i} = P_{k,i} - b_{k,i} P_{k,i} \varphi_{k,i} \varphi_{k,i}^T P_{k,i}, \quad (10)$$

$$b_{k,i} = (1 + \varphi_{k,i}^T P_{k,i} \varphi_{k,i})^{-1}, \quad (11)$$

- 2: Combine (generate $P_{k+1,i}^{-1}$ and $\theta_{k+1,i}$ by a convex combination of $\bar{P}_{k+1,j}^{-1}$ and $\bar{\theta}_{k+1,j}$):

$$P_{k+1,i}^{-1} = \sum_{j \in \mathcal{N}_i} a_{ji} \bar{P}_{k+1,j}^{-1}, \quad (12)$$

$$\theta_{k+1,i} = P_{k+1,i} \sum_{j \in \mathcal{N}_i} a_{ji} \bar{P}_{k+1,j}^{-1} \bar{\theta}_{k+1,j}. \quad (13)$$

When $\mathcal{A} = I_n$, the above distributed LS will degenerate to the classical LS. One may also perform the combination stage for more steps to improve the performance of the algorithm, see e.g. [13]. Note that the diffusion strategy used above is called the covariance intersection fusion rule in (e.g., [46], [47]), and that the above distributed LS algorithm can be deduced from the distributed KF algorithms [13]–[15] by assuming that the state to be estimated is a constant parameter. In this paper, we are interested in the case where each sensor in the network expects to estimate the unknown parameter for its

decision, which is a problem widely studied in the literature, see e.g., [1]–[29]. Here we focus on the scenario where the individual sensor has insufficient information and capability to fulfill the estimation task. It is well known that the estimates and covariance matrices from different sensors may contain complementary information. Combining these two kinds of information together may help to achieve a more accurate estimation of the unknown parameter. Moreover, as stated in [13], the unaware reuse of the same data due to the presence of loops within the network as well as the possible correlation between measurements of different sensors may lead to inconsistency and divergence, which is the primary motivation that leads to the development of the so-called covariance intersection fusion rule [46], [47]. Thus, in order to guarantee the convergence of the estimates for non-independent signals, sometimes it may not be sufficient enough to only exchange information about the estimates.

Note that in the above distributed LS, the computation complexity of each sensor is $O(m^3)$. Moreover, every sensor needs to communicate a total of $(m^2 + 3m)/2$ scalars to its neighboring nodes, and to store a total of $2m^2 + 5m + n + 2$ scalars locally at each time instant k . The algorithm is going to be time-consuming when m is very large, and the covariance intersection fusion rule would only be beneficial when the number of the parameters is manageable locally. Note that if the matrix $\bar{P}_{k,i}$ degenerates to a scalar, for examples, in stochastic gradient-base [41] and LMS-based [30]–[33] distributed estimation algorithms, the communication complexity will be reduced. However, for those algorithms, the estimation error either converges slowly to zero or does not converge to zero at all. Therefore, there is a tradeoff between the complexity and the convergence rate of the distributed estimation algorithms. Moreover, the convergence rate would be “optimal” when $\bar{P}_{k,i}$ is chosen to be the form in the paper. Furthermore, some existing methods can be used to reduce the communication complexity and to make the algorithm suitable for higher dimensional signals, for examples, event-driven methods [55], partial diffusion methods [21], [26], [27], and compressed methods [56] and so on.

III. THE MAIN RESULTS

A. Some Preliminaries

For the theoretical analysis, we need the following standard condition on the noise processes.

Condition 3.1 (Noise condition). For any $i \in \{1, \dots, n\}$, the noise sequence $\{w_{k,i}, \mathcal{F}_k\}$ is a martingale difference (where $\{\mathcal{F}_k\}$ is a sequence of nondecreasing σ -algebras), and there exists a constant $\beta > 2$ such that

$$\sup_{k \geq 0} \mathbb{E}[|w_{k+1,i}|^\beta | \mathcal{F}_k] < \infty, \quad a.s. \quad (14)$$

In order to guarantee the convergence of the above distributed LS algorithm, the following condition on the network topology is naturally required to avoid isolated nodes in the network.

Condition 3.2 (Network topology). The graph \mathcal{G} is connected.

Remark 3.1. From *Lemma 8.1.2* in [57], it is not difficult to see that for any two nodes i and j , there exists a path from i to j with length not less than ℓ if and only if the (i, j) th entry of the matrix \mathcal{A}^ℓ is positive. From this, it is easy to see that each entry of the matrix \mathcal{A}^ℓ will be positive when ℓ is not smaller than the diameter of the graph \mathcal{G} , i.e., $D_{\mathcal{G}}$, see also [15].

B. Theoretical Results

For convenience of analysis, we need to introduce the following notations:

$$\begin{aligned}
\mathbf{Y}_{k+1} &\triangleq \text{col}\{y_{k+1,1}, \dots, y_{k+1,n}\}, & (n \times 1) \\
\Phi_k &\triangleq \text{diag}\{\varphi_{k,1}, \dots, \varphi_{k,n}\}, & (mn \times n) \\
\mathbf{W}_{k+1} &\triangleq \text{col}\{w_{k+1,1}, \dots, w_{k+1,n}\}, & (n \times 1) \\
\Theta &\triangleq \text{col}\{\underbrace{\boldsymbol{\theta}, \dots, \boldsymbol{\theta}}_n\}, & (mn \times 1) \\
\Theta_k &\triangleq \text{col}\{\boldsymbol{\theta}_{k,1}, \dots, \boldsymbol{\theta}_{k,n}\}, & (mn \times 1) \\
\bar{\Theta}_k &\triangleq \text{col}\{\bar{\boldsymbol{\theta}}_{k,1}, \dots, \bar{\boldsymbol{\theta}}_{k,n}\}, & (mn \times 1) \\
\tilde{\Theta}_k &\triangleq \text{col}\{\tilde{\boldsymbol{\theta}}_{k,1}, \dots, \tilde{\boldsymbol{\theta}}_{k,n}\}, & (mn \times 1) \\
&\quad \text{where } \tilde{\boldsymbol{\theta}}_{k,i} = \boldsymbol{\theta} - \boldsymbol{\theta}_{k,i}, \\
\tilde{\bar{\Theta}}_k &\triangleq \text{col}\{\tilde{\bar{\boldsymbol{\theta}}}_{k,1}, \dots, \tilde{\bar{\boldsymbol{\theta}}}_{k,n}\}, & (mn \times 1) \\
&\quad \text{where } \tilde{\bar{\boldsymbol{\theta}}}_{k,i} = \boldsymbol{\theta} - \bar{\boldsymbol{\theta}}_{k,i}, \\
\mathbf{b}_k &\triangleq \text{diag}\{b_{k,1}, \dots, b_{k,n}\}, & (n \times n) \\
\mathbf{c}_k &\triangleq \mathbf{b}_k \otimes I_m, & (mn \times mn) \\
\mathbf{P}_k &\triangleq \text{diag}\{P_{k,1}, \dots, P_{k,n}\}, & (mn \times mn) \\
\bar{\mathbf{P}}_k &\triangleq \text{diag}\{\bar{P}_{k,1}, \dots, \bar{P}_{k,n}\}, & (mn \times mn) \\
\mathcal{A} &\triangleq \mathcal{A} \otimes I_m, & (mn \times mn)
\end{aligned}$$

where $\text{col}\{\dots\}$ denotes a vector by stacking the specified vectors, $\text{diag}\{\dots\}$ is used in a non-standard manner which means that $m \times 1$ column vectors are combined ‘‘in a diagonal manner’’ resulting in a $mn \times n$ matrix, and \otimes is the Kronecker product. Note also that Θ is just the n -times replication of vectors $\boldsymbol{\theta}$, and the matrix \mathcal{A} is the weighted adjacency matrix of the graph \mathcal{G} .

Then (1) can be rewritten in the following compact form:

$$\mathbf{Y}_{k+1} = \Phi_k^T \Theta + \mathbf{W}_{k+1}, \quad (15)$$

Similarly, for the distributed LS algorithm we have

$$\left\{ \begin{array}{l}
\bar{\Theta}_{k+1} = \Theta_k + \mathbf{c}_k \mathbf{P}_k \Phi_k (\mathbf{Y}_{k+1} - \Phi_k^T \Theta_k), \\
\bar{\mathbf{P}}_{k+1} = \mathbf{P}_k - \mathbf{c}_k \mathbf{P}_k \Phi_k \Phi_k^T \mathbf{P}_k, \\
\mathbf{b}_k = (I_n + \Phi_k^T \mathbf{P}_k \Phi_k)^{-1}, \\
\mathbf{c}_k = \mathbf{b}_k \otimes I_m, \\
\text{vec}\{\mathbf{P}_{k+1}^{-1}\} = \mathcal{A} \text{vec}\{\bar{\mathbf{P}}_{k+1}^{-1}\}, \\
\Theta_{k+1} = \mathbf{P}_{k+1} \mathcal{A} \bar{\mathbf{P}}_{k+1}^{-1} \bar{\Theta}_{k+1},
\end{array} \right. \quad (16)$$

where $\text{vec}\{\cdot\}$ denotes the operator that stacks the blocks of a block diagonal matrix on top of each other.

Since $\tilde{\Theta}_k = \Theta - \Theta_k$ and $\tilde{\bar{\Theta}}_k = \Theta - \bar{\Theta}_k$ by definition, substituting (15) into (16), we can get

$$\begin{aligned}
\tilde{\Theta}_{k+1} &= \Theta - \bar{\Theta}_{k+1} \\
&= \Theta - \Theta_k - \mathbf{c}_k \mathbf{P}_k \Phi_k (\Phi_k^T \Theta + \mathbf{W}_{k+1} - \Phi_k^T \Theta_k) \\
&= (I_{mn} - \mathbf{c}_k \mathbf{P}_k \Phi_k \Phi_k^T) \tilde{\Theta}_k - \mathbf{c}_k \mathbf{P}_k \Phi_k \mathbf{W}_{k+1} \\
&= \bar{\mathbf{P}}_{k+1} \mathbf{P}_k^{-1} \tilde{\Theta}_k - \mathbf{c}_k \mathbf{P}_k \Phi_k \mathbf{W}_{k+1}.
\end{aligned}$$

Note also that

$$\begin{aligned}
&\mathbf{P}_{k+1} \mathcal{A} \bar{\mathbf{P}}_{k+1}^{-1} \Theta \\
&= \text{col} \left\{ P_{k+1,1} \sum_{j \in \mathcal{N}_1} a_{j1} \bar{P}_{k+1,j}^{-1} \boldsymbol{\theta}, \dots, P_{k+1,n} \sum_{j \in \mathcal{N}_n} a_{jn} \bar{P}_{k+1,j}^{-1} \boldsymbol{\theta} \right\}.
\end{aligned}$$

Then for each sensor $i \in \{1, 2, \dots, n\}$,

$$P_{k+1,i} \sum_{j \in \mathcal{N}_i} a_{ji} \bar{P}_{k+1,j}^{-1} \boldsymbol{\theta} = \left[P_{k+1,i} \left(\sum_{j \in \mathcal{N}_i} a_{ji} \bar{P}_{k+1,j}^{-1} \right) \right] \boldsymbol{\theta} = \boldsymbol{\theta}.$$

Thus, $\Theta = \mathbf{P}_{k+1} \mathcal{A} \bar{\mathbf{P}}_{k+1}^{-1} \Theta$ holds. Then we have

$$\begin{aligned}
\tilde{\Theta}_{k+1} &= \Theta - \bar{\Theta}_{k+1} \\
&= \Theta - \mathbf{P}_{k+1} \mathcal{A} \bar{\mathbf{P}}_{k+1}^{-1} \bar{\Theta}_{k+1} \\
&= \mathbf{P}_{k+1} \mathcal{A} \bar{\mathbf{P}}_{k+1}^{-1} \Theta - \mathbf{P}_{k+1} \mathcal{A} \bar{\mathbf{P}}_{k+1}^{-1} \bar{\Theta}_{k+1} \\
&= \mathbf{P}_{k+1} \mathcal{A} \bar{\mathbf{P}}_{k+1}^{-1} \tilde{\Theta}_{k+1} \\
&= \mathbf{P}_{k+1} \mathcal{A} \mathbf{P}_k^{-1} \tilde{\Theta}_k \\
&\quad - \mathbf{P}_{k+1} \mathcal{A} \bar{\mathbf{P}}_{k+1}^{-1} \mathbf{c}_k \mathbf{P}_k \Phi_k \mathbf{W}_{k+1}. \quad (17)
\end{aligned}$$

Before establishing a theory on the learning and prediction behavior of the distributed LS, we first present a critical theorem, which requires no excitation conditions on the regression process $\varphi_{k,i}$.

Theorem 3.1. Let *Condition 3.1* be satisfied, we have as $t \rightarrow \infty$,

$$1) \sum_{k=0}^t \tilde{\Theta}_k^T \Phi_k \mathbf{b}_k \Phi_k^T \tilde{\Theta}_k = O(\log(r_t)), \quad a.s.,$$

$$2) \tilde{\Theta}_{t+1}^T \mathbf{P}_{t+1}^{-1} \tilde{\Theta}_{t+1} = O(\log(r_t)), \quad a.s.,$$

where

$$r_t = \lambda_{\max}\{\mathbf{P}_0^{-1}\} + \sum_{i=1}^n \sum_{k=0}^t \|\varphi_{k,i}\|^2. \quad (18)$$

The detailed proof of *Theorem 3.1* is supplied in the next section. From this, we can obtain the following upper bound of the accumulated regrets for the distributed LS-based adaptive predictor.

Theorem 3.2. Let *Condition 3.1* be satisfied. Then the sample paths of the accumulated regrets have the following bound as $t \rightarrow \infty$:

$$\sum_{i=1}^n \sum_{k=0}^t R_{k,i} = O(\log(r_t)), \quad a.s., \quad (19)$$

provided that $\Phi_t^T \mathbf{P}_t \Phi_t = O(1)$, *a.s.*

The proof of *Theorems 3.2* is given in Section IV.

Remark 3.2. We remark that when the regressors at each node are bounded in the time-averaging sense, then r_t will be of the order $O(t)$, and consequently by *Theorem 3.2*, we know

that the bound on the accumulated regret (19) will be sublinear with respect to nt , i.e., $\frac{1}{nt} \sum_{i=1}^n \sum_{k=0}^t R_{k,i} = O(\frac{\log t}{t}) \rightarrow 0$, as $t \rightarrow \infty$, i.e., the averaged regret goes to zero and the distributed LS algorithm for the prediction problem performs well. The order $O(\log(r_t))$ for the accumulated regrets may be shown to be the best possible among all adaptive predictors, as is already known in the traditional single sensor case, see [58]. The precise constant in $O(\cdot)$ may also be determined if we have further conditions on the regressors, see *Corollary 3.3* in [33] in the single sensor case.

We point out that one can also get precise upper bound for the expected accumulated regrets for any finite $t \geq 1$, which is stated in the following remark.

Remark 3.3. Let *Condition 3.1* be satisfied. Then the expected accumulated regrets have the following bound for any $t \geq 1$:

$$\sum_{i=1}^n \sum_{k=0}^t \mathbb{E}[R_{k,i}] \leq a \log(\mathbb{E}[r_t]) + b,$$

provided that $\mathbb{E}[\|\varphi_{k,i}\|^2] < \infty, \forall k \geq 0, \forall i \in \{1, \dots, n\}$, and there exists deterministic constants $c > 0, \bar{\sigma} > 0$ such that $\|\Phi_t^T P_t \Phi_t\| \leq c, \sigma_w \leq \bar{\sigma}$, where

$$\begin{aligned} a &= (1+c)mn\bar{\sigma}, \\ b &= (1+c) \left\{ \mathbb{E}[\tilde{\Theta}_0^T P_0^{-1} \tilde{\Theta}_0] - \bar{\sigma} \mathbb{E}[\log(|P_0^{-1}|)] \right\}. \end{aligned}$$

Note that

$$\sigma_w \triangleq \sum_{i=1}^n \sigma_i^2, \quad \sigma_i^2 \triangleq \sup_{k \geq 0} \mathbb{E}[w_{k+1,i}^2 | \mathcal{F}_k], \quad (20)$$

which is finite almost surely by *Condition 3.1*. The detailed proof is given in Appendix B.

From *Theorem 3.1*, we can also obtain the strong consistency of the distributed LS to guarantee the generalization ability of learning, under the following cooperative excitation condition.

Condition 3.3 (Cooperative excitation condition). The growth rate of $\log(\lambda_{\max}\{P_k^{-1}\})$ is slower than that of $\lambda_{\min}\{P_k^{-1}\}$, in other words,

$$\lim_{t \rightarrow \infty} \frac{\log(r_t)}{\lambda_{\min}^{n,t}} = 0, \quad a.s., \quad (21)$$

where r_t is defined by (18), and

$$\lambda_{\min}^{n,t} = \lambda_{\min} \left\{ \sum_{j=1}^n P_{0,j}^{-1} + \sum_{j=1}^n \sum_{k=0}^{t-D_G+1} \varphi_{k,j} \varphi_{k,j}^T \right\}.$$

Remark 3.4. Let us give some intuitive explanations for *Condition 3.3* used in the paper. To start with, let us first consider the extreme case where the regressor process φ_j^i is identically zero. It is clear that *Condition 3.3* is not satisfied, which is indeed a trivial case where the system is not identifiable since the observations contain no information about the unknown parameters. Hence, to estimate the unknown parameters, some non-zero ‘‘excitation’’ conditions should be imposed on the regressors φ_j^i , which are usually reflected in the so called (Fisher) information matrix P_k^{-1} , and now explicitly required in our *Condition 3.3*. We remark that in

the traditional single sensor case (where $n = 1$ and $D_G = 1$), *Condition 3.3* reduces to the well-known Lai-Wei excitation condition (8) for $i = 1$, which is known to be the weakest possible data condition for the convergence of the classical LS estimates [38]. This condition is much weaker than the well-known persistence of excitation (PE) condition usually used in the parameter estimation of finite-dimensional linear control systems, since the PE condition requires that the condition number of P_k^{-1} , i.e., $\frac{\lambda_{\max}\{P_k^{-1}\}}{\lambda_{\min}\{P_k^{-1}\}}$ is bounded a.s. for all $k \geq 1$. Moreover, it is easy to convince oneself that the cooperative excitation condition (*Condition 3.3*) will make it possible for the distributed LS to consistently estimate the unknown parameter, even if any individual sensor cannot due to lack of suitable excitation, e.g., when (8) is not satisfied, because *Condition 3.3* is obviously weaker than (8) for any i . Finally, we remark that the verification of *Condition 3.3* is straightforward in the ergodic case, since $\log(r_t)$ is of the order $O(\log t)$, and $\lambda_{\min}^{n,t}/t$ tends to $\lambda_{\min}\{\sum_{j=1}^n \mathbb{E}[\varphi_{0,j} \varphi_{0,j}^T]\}$ as $t \rightarrow \infty$, which will be positive if the expectation of the summation of the covariance matrices is positive definite. For more general correlated non-stationary signals from control systems, the verification of *Condition 3.3* may be conducted following a similar way as that for the traditional single sensor case (see, [41]).

Theorem 3.3 below states that if *Condition 3.3* holds, then the distributed LS estimate Θ_t will converge to the true unknown parameter.

Theorem 3.3. Let *Conditions 3.1* and *3.2* be satisfied, we have as $t \rightarrow \infty$,

$$\|\tilde{\Theta}_{t+1}\|^2 = O\left(\frac{\log(r_t)}{\lambda_{\min}^{n,t}}\right), \quad a.s., \quad (22)$$

where r_t is defined by (18) and $\lambda_{\min}^{n,t}$ is defined in *Condition 3.3*.

Remark 3.5. The detailed proof of *Theorem 3.3* is given in the next section. We remark that the upper bound of the estimation error $\tilde{\Theta}_{t+1}$ established in *Theorem 3.3* does not need *Condition 3.3*. It is needed only when the estimation error $\tilde{\Theta}_{t+1}$ is required to approach zero. Moreover, the above theoretical analysis method can naturally be generalized to multidimensional cases, e.g. the widely used autoregressive-moving average with exogenous input (ARMAX) model [41], where the unknown parameter is a matrix, both the regressors and observations are stochastic vectors, and the noises are colored.

Note that the linear stochastic regression model (1) is a basic hypothesis for our theoretical investigation, which can be regarded as an approximation of more complex systems and is widely used and studied in many different fields, e.g., automatic control, signal processing, statistics, adaptive filtering, distributed estimation, and so on. Note also that the linearity in the model (1) is only assumed for the unknown parameter θ , it can be nonlinear in terms of the input and output data in the regressor $\varphi_{k,i}$. Of course, when the data does not satisfy such a model, the estimates may be biased and the problem as well as the corresponding theory should be reformulated and investigated. If we assume that the noise process $w_{k+1,i}$ contains not only

the observation noise satisfying *Condition 3.1*, but also some unknown dynamics (or model bias) which is assumed to be bounded, then it is not difficult to prove that the above regret bound will depend on the bound of the unknown dynamics under the PE condition [41]. Moreover, if we assume that the observation model contains some types of bias, then the deviation in the estimates may either be corrected by some bias-compensation techniques [25], or be approximated by using a regression model with slowly increasing lags (see [41], Chapter 9). Furthermore, some model validation methods may be also used to estimate the bound of the unknown dynamics (or model bias) when the ideal mathematical model is biased [59]. In all these cases, the analyses in this paper should serve as a basis for further investigation on the related distributed estimation problems.

Remark 3.6. Let us now compare the above distributed LS algorithm with centralized methods whereby, at each time instant k , all the n sensors transmit their raw data $\{y_{k+1,i}, \varphi_{k,i}\}$ to a fusion center for processing to obtain the centralized estimate θ_{k+1}^c . Note that there are many different ways to construct a centralized algorithm, which may give different estimation errors. Let us consider a simple and natural way in the following. Denote

$$\begin{aligned} \mathbf{Y}_{k+1} &\triangleq \text{col}\{y_{k+1,1}, \dots, y_{k+1,n}\}, & (n \times 1) \\ \mathbf{W}_{k+1} &\triangleq \text{col}\{w_{k+1,1}, \dots, w_{k+1,n}\}, & (n \times 1) \\ \Phi_k^c &\triangleq (\varphi_{k,1}, \dots, \varphi_{k,n}), & (m \times n) \end{aligned}$$

then one has the following regression model:

$$\mathbf{Y}_{k+1} = (\Phi_k^c)^T \theta + \mathbf{W}_{k+1}.$$

Let the centralized LS estimate be defined by the following at each time instant k :

$$\theta_k^c = \arg \min_{\theta \in \mathbb{R}^m} \sum_{j=1}^k [\mathbf{Y}_j - (\Phi_{j-1}^c)^T \theta]^T [\mathbf{Y}_j - (\Phi_{j-1}^c)^T \theta],$$

which can be calculated recursively as follows:

$$\begin{aligned} \theta_{k+1}^c &= \theta_k^c + P_k \Phi_k^c B_k [\mathbf{Y}_{k+1} - (\Phi_k^c)^T \theta_k^c], \\ P_{k+1} &= P_k - P_k \Phi_k^c B_k (\Phi_k^c)^T P_k, \\ B_k &= [I_n + (\Phi_k^c)^T P_k \Phi_k^c]^{-1}, \end{aligned}$$

where the initial estimate $\theta_0^c \in \mathbb{R}^m$, and the initial positive definite matrix $P_0 \in \mathbb{R}^{m \times m}$ can be chosen arbitrarily. Then by (7), the above centralized LS has the following upper bound for the estimation error as $k \rightarrow \infty$:

$$\begin{aligned} &\|\theta_{k+1}^c - \theta\|^2 \\ &= O\left(\frac{\log(\lambda_{\max}\{P_0^{-1}\} + \sum_{j=0}^k \|\Phi_j^c\|^2)}{\lambda_{\min}\{P_0^{-1} + \sum_{j=0}^k \Phi_j^c (\Phi_j^c)^T\}}\right), \quad a.s. \\ &= O\left(\frac{\log(\lambda_{\max}\{P_0^{-1}\} + \sum_{i=1}^n \sum_{j=0}^k \|\varphi_{j,i}\|^2)}{\lambda_{\min}\{P_0^{-1} + \sum_{i=1}^n \sum_{j=0}^k \varphi_{j,i} (\varphi_{j,i})^T\}}\right), \quad a.s. \end{aligned}$$

From this and *Theorem 3.3* one can see that both the convergence condition and the convergence rate of the cen-

tralized algorithm is essentially the same as those for the distributed algorithm. Moreover, for the centralized algorithm, the computation complexity of the fusion center is $O(m^3 + m^2n + mn^2 + n^3)$, which is of the same order compared with the computation complexity of **Algorithm 1**. Every sensor needs to communicate a total of $m+1$ scalars to the fusion center, and the fusion center needs to communicate a total of m scalars to each sensor and store a total of $(m^2 + 3m + n^2 + 3n)/2 + mn$ scalars at each time instant k . Although the centralized algorithm has some advantages over the distributed algorithm in terms of communication complexity, it also has some drawbacks compared with the distributed case. Firstly, the distributed methods may have stronger structural robustness compared with the centralized ones. This is because the centralized algorithm will fail once the fusion center is broken down by outside attacks, while the distributed algorithm can still estimate the unknown parameters even if the communications among some sensors are interrupted, as long as the network connectivity is maintained. Secondly, if the fusion center is far away from some sensors, the communications with the fusion center may not be feasible, and the transmission of observations and regression vectors may compromise the safety and privacy of the system even if the communication is possible. Hence, there may be many factors need to be considered when we choose to use the centralized or distributed algorithms.

IV. PROOFS OF THE MAIN RESULTS

A. Proof of Theorem 3.1

To prove *Theorem 3.1*, we need to establish several lemmas first. The first lemma below is a key inequality on convex combination of nonnegative definite matrices.

Lemma 4.1. For any adjacency matrix $\mathcal{A} = \{a_{ij}\} \in \mathbb{R}^{n \times n}$, denote $\mathcal{A} = \mathcal{A} \otimes I_m$, and for any nonnegative definite matrices $Q_i \in \mathbb{R}^{m \times m}$, $i = 1, \dots, n$, denote

$$\begin{aligned} Q &= \text{diag}\{Q_1, \dots, Q_n\}, \\ Q' &= \text{diag}\{Q'_1, \dots, Q'_n\}, \end{aligned}$$

where $Q'_i = \sum_{j=1}^n a_{ji} Q_j$. Then the following inequality holds:

$$\mathcal{A} Q \mathcal{A} \leq Q'. \quad (23)$$

Proof: By the definition of \mathcal{A} and Q , we can get that

$$\mathcal{A} Q \mathcal{A} = \begin{pmatrix} \sum_{j=1}^n a_{1j} a_{j1} Q_j & \cdots & \sum_{j=1}^n a_{1j} a_{jn} Q_j \\ \sum_{j=1}^n a_{2j} a_{j1} Q_j & \cdots & \sum_{j=1}^n a_{2j} a_{jn} Q_j \\ \vdots & \ddots & \vdots \\ \sum_{j=1}^n a_{nj} a_{j1} Q_j & \cdots & \sum_{j=1}^n a_{nj} a_{jn} Q_j \end{pmatrix}.$$

In order to prove (23), we only need to prove that for any unit column vector $x \in \mathbb{R}^{mn}$ with $\|x\| = 1$, $x^T \mathcal{A} Q \mathcal{A} x \leq x^T Q' x$ holds. Denote $x = \text{col}\{x_1, x_2, \dots, x_n\}$ with $x_i \in \mathbb{R}^m$,

then by the Schwarz inequality and noticing that $Q_j \geq 0$, $\sum_{j=1}^n a_{ij} = 1$, and $a_{ji} = a_{ij}$, ($i, j = 1, \dots, n$), we have

$$\begin{aligned}
& x^T \mathcal{A} Q \mathcal{A} x \\
&= \sum_{p=1}^n \sum_{q=1}^n \sum_{j=1}^n a_{pj} a_{jq} x_p^T Q_j x_q \\
&= \sum_{p=1}^n \sum_{q=1}^n \sum_{j=1}^n \sqrt{a_{pj} a_{jq}} x_p^T Q_j^{\frac{1}{2}} \cdot \sqrt{a_{pj} a_{jq}} Q_j^{\frac{1}{2}} x_q \\
&\leq \left\{ \sum_{p=1}^n \sum_{q=1}^n \sum_{j=1}^n a_{pj} a_{jq} x_p^T Q_j x_p \right\}^{\frac{1}{2}} \\
&\quad \cdot \left\{ \sum_{p=1}^n \sum_{q=1}^n \sum_{j=1}^n a_{pj} a_{jq} x_q^T Q_j x_q \right\}^{\frac{1}{2}} \\
&= \left\{ \sum_{p=1}^n \sum_{j=1}^n a_{pj} x_p^T Q_j x_p \right\}^{\frac{1}{2}} \left\{ \sum_{q=1}^n \sum_{j=1}^n a_{jq} x_q^T Q_j x_q \right\}^{\frac{1}{2}} \\
&= \sum_{i=1}^n \sum_{j=1}^n a_{ji} x_i^T Q_j x_i \\
&= x^T Q' x,
\end{aligned}$$

which completes the proof.

By Lemma 4.1, we can obtain the following result.

Lemma 4.2. For any adjacency matrix $\mathcal{A} = \{a_{ij}\} \in \mathbb{R}^{n \times n}$, denote $\mathcal{A} = \mathcal{A} \otimes I_m$. Then for any $k \geq 1$,

$$\mathcal{A} \bar{P}_{k+1}^{-1} \mathcal{A} \leq P_{k+1}^{-1}, \quad (24)$$

and

$$\mathcal{A} P_{k+1} \mathcal{A} \leq \bar{P}_{k+1}, \quad (25)$$

holds, where \bar{P}_{k+1} and P_{k+1} are defined in (16).

Proof: By taking $Q_i = \bar{P}_{k+1,i}^{-1} \geq 0$ and noticing $P_{k+1,i}^{-1} = \sum_{j=1}^n a_{ji} \bar{P}_{k+1,j}^{-1} = Q'_i$, we know from Lemma 4.1 that

$$\mathcal{A} \bar{P}_{k+1}^{-1} \mathcal{A} \leq P_{k+1}^{-1},$$

holds. To prove (25), we first assume that \mathcal{A} is invertible. Then by Lemma A.1 in Appendix A, it is easy to see that

$$\mathcal{A} P_{k+1} \mathcal{A} \leq \bar{P}_{k+1}.$$

Next, we consider the case where \mathcal{A} is not invertible. Since the number of eigenvalues of the matrix \mathcal{A} is finite, then exists a constant $\varepsilon^* \in (0, 1)$ such that the perturbed adjacency matrix $\mathcal{A}^\varepsilon = \mathcal{A} + \varepsilon I_{mn} = \{a_{ij}^\varepsilon\}$ will be invertible for any $0 < \varepsilon < \varepsilon^*$. By the definition of \mathcal{A}^ε , we know that \mathcal{A}^ε is symmetric and the sums of each columns and rows of the matrix \mathcal{A}^ε are all $1 + \varepsilon$. Then we define

$$(P_{k+1,i}^\varepsilon)^{-1} = \sum_{j=1}^n a_{ji}^\varepsilon \bar{P}_{k+1,j}^{-1},$$

and we can denote $P_{k+1}^\varepsilon = \text{diag}\{P_{k+1,1}^\varepsilon, \dots, P_{k+1,n}^\varepsilon\}$ since $(P_{k+1,i}^\varepsilon)^{-1}$ defined above is invertible. Similar to the proof of

Lemma 4.1, for any unit column vector $x \in \mathbb{R}^{mn}$, we have

$$\begin{aligned}
& x^T \mathcal{A}^\varepsilon \bar{P}_{k+1}^{-1} \mathcal{A}^\varepsilon x \\
&\leq \left\{ \sum_{p=1}^n \sum_{q=1}^n \sum_{j=1}^n a_{pj}^\varepsilon a_{jq}^\varepsilon x_p^T \bar{P}_{k+1,j}^{-1} x_p \right\}^{\frac{1}{2}} \\
&\quad \cdot \left\{ \sum_{p=1}^n \sum_{q=1}^n \sum_{j=1}^n a_{pj}^\varepsilon a_{jq}^\varepsilon x_q^T \bar{P}_{k+1,j}^{-1} x_q \right\}^{\frac{1}{2}} \\
&= (1 + \varepsilon) \sum_{i=1}^n \sum_{j=1}^n a_{ji}^\varepsilon x_i^T \bar{P}_{k+1,j}^{-1} x_i \\
&= (1 + \varepsilon) x^T (P_{k+1}^\varepsilon)^{-1} x.
\end{aligned}$$

Consequently, we have $\mathcal{A}^\varepsilon \bar{P}_{k+1}^{-1} \mathcal{A}^\varepsilon \leq (1 + \varepsilon) (P_{k+1}^\varepsilon)^{-1}$. Since \mathcal{A}^ε is invertible, we know from Lemma A.1 in Appendix A that

$$\mathcal{A}^\varepsilon P_{k+1}^\varepsilon \mathcal{A}^\varepsilon \leq (1 + \varepsilon) \bar{P}_{k+1}.$$

By taking $\varepsilon \rightarrow 0$ on both sides of the above equation, we can obtain that

$$\lim_{\varepsilon \rightarrow 0} \mathcal{A}^\varepsilon P_{k+1}^\varepsilon \mathcal{A}^\varepsilon = \mathcal{A} P_{k+1} \mathcal{A} \leq \lim_{\varepsilon \rightarrow 0} (1 + \varepsilon) \bar{P}_{k+1} = \bar{P}_{k+1}.$$

This completes the proof. \blacksquare

To accomplish the proof of Theorem 3.1, we also need the following inequality.

Lemma 4.3. For any adjacency matrix $\mathcal{A} = \{a_{ij}\} \in \mathbb{R}^{n \times n}$, and for any $k \geq 1$,

$$|\bar{P}_{k+1}^{-1}| \leq |P_{k+1}^{-1}|, \quad (26)$$

holds, where \bar{P}_{k+1} and P_{k+1} are defined in (16).

Proof: Since

$$P_{k+1}^{-1} = \begin{pmatrix} \sum_{j=1}^n a_{j1} \bar{P}_{k+1,j}^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sum_{j=1}^n a_{jn} \bar{P}_{k+1,j}^{-1} \end{pmatrix},$$

and

$$\bar{P}_{k+1}^{-1} = \begin{pmatrix} \bar{P}_{k+1,1}^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \bar{P}_{k+1,n}^{-1} \end{pmatrix},$$

by Lemma A.2 in Appendix A and noticing the definition of the adjacency matrix $\mathcal{A} = \{a_{ij}\}$, we can see that

$$\begin{aligned}
|P_{k+1}^{-1}| &= \prod_{i=1}^n \left| \sum_{j=1}^n a_{ji} \bar{P}_{k+1,j}^{-1} \right| \\
&\geq \prod_{i=1}^n |\bar{P}_{k+1,1}^{-1}|^{a_{1i}} |\bar{P}_{k+1,2}^{-1}|^{a_{2i}} \cdots |\bar{P}_{k+1,n}^{-1}|^{a_{ni}} \\
&= |\bar{P}_{k+1,1}^{-1}|^{\sum_{i=1}^n a_{1i}} |\bar{P}_{k+1,2}^{-1}|^{\sum_{i=1}^n a_{2i}} \cdots |\bar{P}_{k+1,n}^{-1}|^{\sum_{i=1}^n a_{ni}} \\
&= |\bar{P}_{k+1,1}^{-1}| \cdot |\bar{P}_{k+1,2}^{-1}| \cdots |\bar{P}_{k+1,n}^{-1}| \\
&= |\bar{P}_{k+1}^{-1}|,
\end{aligned}$$

which completes the proof. \blacksquare

To prove Theorem 3.1, we also need the following critical

lemma.

Lemma 4.4. Let *Condition 3.1* be satisfied. Then the distributed LS defined by (15) and (16) satisfies the following relationship as $t \rightarrow \infty$:

$$\begin{aligned} & \tilde{\Theta}_{t+1}^T P_{t+1}^{-1} \tilde{\Theta}_{t+1} \\ & + [1 + o(1)] \sum_{k=0}^t \tilde{\Theta}_k^T \Phi_k \mathbf{b}_k \Phi_k^T \tilde{\Theta}_k \\ & + [1 + o(1)] \sum_{k=0}^t \tilde{\Theta}_k^T P_k^{-1} \Delta_{k+1} P_k^{-1} \tilde{\Theta}_k \\ & \leq \sigma_w \log(|P_{t+1}^{-1}|) + o(\log(|P_{t+1}^{-1}|)) + O(1), \quad a.s., \quad (27) \end{aligned}$$

where $\mathbf{b}_k = (I_n + \Phi_k^T P_k \Phi_k)^{-1}$, $\mathbf{c}_k = \mathbf{b}_k \otimes I_m$, $\Delta_{k+1} \triangleq \bar{P}_{k+1} - \mathcal{A} P_{k+1} \mathcal{A} \geq 0$ by *Lemma 4.2*, and σ_w is defined by (20).

Proof: Since $\mathbf{b}_k = (I_n + \Phi_k^T P_k \Phi_k)^{-1}$ and $\mathbf{c}_k = \mathbf{b}_k \otimes I_m$, then by (17), we know that

$$\tilde{\Theta}_{k+1} = P_{k+1} \mathcal{A} P_k^{-1} \tilde{\Theta}_k - P_{k+1} \mathcal{A} \bar{P}_{k+1}^{-1} \mathbf{c}_k P_k \Phi_k W_{k+1}.$$

Hence, we have the following expansion for the stochastic Lyapunov function $V_k = \tilde{\Theta}_k^T P_k^{-1} \tilde{\Theta}_k$:

$$\begin{aligned} V_{k+1} & = \tilde{\Theta}_{k+1}^T P_{k+1}^{-1} \tilde{\Theta}_{k+1} \\ & = (\tilde{\Theta}_k^T P_k^{-1} \mathcal{A} P_{k+1} - W_{k+1}^T \Phi_k^T P_k \mathbf{c}_k \bar{P}_{k+1}^{-1} \mathcal{A} P_{k+1}) \\ & \quad \cdot (\mathcal{A} P_k^{-1} \tilde{\Theta}_k - \mathcal{A} \bar{P}_{k+1}^{-1} \mathbf{c}_k P_k \Phi_k W_{k+1}) \\ & = \tilde{\Theta}_k^T P_k^{-1} \mathcal{A} P_{k+1} \mathcal{A} P_k^{-1} \tilde{\Theta}_k \\ & \quad - 2 \tilde{\Theta}_k^T P_k^{-1} \mathcal{A} P_{k+1} \mathcal{A} \bar{P}_{k+1}^{-1} \mathbf{c}_k P_k \Phi_k W_{k+1} \\ & \quad + W_{k+1}^T \Phi_k^T P_k \mathbf{c}_k \bar{P}_{k+1}^{-1} \mathcal{A} P_{k+1} \mathcal{A} \bar{P}_{k+1}^{-1} \\ & \quad \cdot \mathbf{c}_k P_k \Phi_k W_{k+1}. \quad (28) \end{aligned}$$

Now, we proceed to estimate the right-hand-side (RHS) of (28) term by term. Firstly, we know that

$$\begin{aligned} & \tilde{\Theta}_k^T P_k^{-1} \mathcal{A} P_{k+1} \mathcal{A} P_k^{-1} \tilde{\Theta}_k \\ & = \tilde{\Theta}_k^T P_k^{-1} \bar{P}_{k+1} P_k^{-1} \tilde{\Theta}_k - \tilde{\Theta}_k^T P_k^{-1} \Delta_{k+1} P_k^{-1} \tilde{\Theta}_k \\ & = \tilde{\Theta}_k^T P_k^{-1} (P_k - P_k \Phi_k \mathbf{b}_k \Phi_k^T P_k) P_k^{-1} \tilde{\Theta}_k \\ & \quad - \tilde{\Theta}_k^T P_k^{-1} \Delta_{k+1} P_k^{-1} \tilde{\Theta}_k \\ & = \tilde{\Theta}_k^T P_k^{-1} \tilde{\Theta}_k - \tilde{\Theta}_k^T \Phi_k \mathbf{b}_k \Phi_k^T \tilde{\Theta}_k \\ & \quad - \tilde{\Theta}_k^T P_k^{-1} \Delta_{k+1} P_k^{-1} \tilde{\Theta}_k \\ & = V_k - \tilde{\Theta}_k^T \Phi_k \mathbf{b}_k \Phi_k^T \tilde{\Theta}_k - \tilde{\Theta}_k^T P_k^{-1} \Delta_{k+1} P_k^{-1} \tilde{\Theta}_k. \quad (29) \end{aligned}$$

Moreover, by the (block) diagonal property of \mathbf{b}_k , \mathbf{c}_k , P_k and Φ_k , we have

$$\mathbf{c}_k P_k = P_k \mathbf{c}_k, \quad \Phi_k^T \mathbf{c}_k = \mathbf{b}_k \Phi_k^T, \quad \mathbf{c}_k \Phi_k = \Phi_k \mathbf{b}_k. \quad (30)$$

By *Lemma A.3* in Appendix A and let $A = P_k^{-1}$, $B = \Phi_k$, $C = \Phi_k^T$ and $D = I_n$ respectively, it is easy to see that

$$\begin{aligned} & (P_k^{-1} + \Phi_k \Phi_k^T)^{-1} \\ & = P_k - P_k \Phi_k (I_n + \Phi_k^T P_k \Phi_k)^{-1} \Phi_k^T P_k \\ & = P_k - P_k \Phi_k \mathbf{b}_k \Phi_k^T P_k \\ & = \bar{P}_{k+1}. \end{aligned}$$

From this, we have $\bar{P}_{k+1}^{-1} = P_k^{-1} + \Phi_k \Phi_k^T$. Thus, we can estimate the second term on the RHS of (28) as follows:

$$\begin{aligned} & \tilde{\Theta}_k^T P_k^{-1} \mathcal{A} P_{k+1} \mathcal{A} \bar{P}_{k+1}^{-1} \mathbf{c}_k P_k \Phi_k W_{k+1} \\ & = \tilde{\Theta}_k^T P_k^{-1} \mathcal{A} P_{k+1} \mathcal{A} (P_k^{-1} + \Phi_k \Phi_k^T) \mathbf{c}_k P_k \Phi_k W_{k+1} \\ & = \tilde{\Theta}_k^T P_k^{-1} \mathcal{A} P_{k+1} \mathcal{A} \mathbf{c}_k \Phi_k W_{k+1} \\ & \quad + \tilde{\Theta}_k^T P_k^{-1} \mathcal{A} P_{k+1} \mathcal{A} \Phi_k \Phi_k^T \mathbf{c}_k P_k \Phi_k W_{k+1} \\ & = \tilde{\Theta}_k^T P_k^{-1} \mathcal{A} P_{k+1} \mathcal{A} \mathbf{c}_k \Phi_k W_{k+1} \\ & \quad + \tilde{\Theta}_k^T P_k^{-1} \mathcal{A} P_{k+1} \mathcal{A} \Phi_k \mathbf{b}_k (I_n + \Phi_k^T P_k \Phi_k) W_{k+1} \\ & \quad - \tilde{\Theta}_k^T P_k^{-1} \mathcal{A} P_{k+1} \mathcal{A} \Phi_k \mathbf{b}_k W_{k+1} \\ & = \tilde{\Theta}_k^T P_k^{-1} \mathcal{A} P_{k+1} \mathcal{A} \mathbf{c}_k \Phi_k W_{k+1} \\ & \quad + \tilde{\Theta}_k^T P_k^{-1} \mathcal{A} P_{k+1} \mathcal{A} \Phi_k W_{k+1} \\ & \quad - \tilde{\Theta}_k^T P_k^{-1} \mathcal{A} P_{k+1} \mathcal{A} \Phi_k \mathbf{b}_k W_{k+1} \\ & = \tilde{\Theta}_k^T P_k^{-1} \mathcal{A} P_{k+1} \mathcal{A} \Phi_k W_{k+1} \\ & = \tilde{\Theta}_k^T P_k^{-1} \bar{P}_{k+1} \Phi_k W_{k+1} - \tilde{\Theta}_k^T P_k^{-1} \Delta_{k+1} \Phi_k W_{k+1}. \quad (31) \end{aligned}$$

As for the last term on the RHS of (28), by $\mathcal{A} P_{k+1} \mathcal{A} \leq \bar{P}_{k+1}$, we can estimate it as follows:

$$\begin{aligned} & W_{k+1}^T \Phi_k^T P_k \mathbf{c}_k \bar{P}_{k+1}^{-1} \mathcal{A} P_{k+1} \mathcal{A} \bar{P}_{k+1}^{-1} \mathbf{c}_k P_k \Phi_k W_{k+1} \\ & \leq W_{k+1}^T \Phi_k^T P_k \mathbf{c}_k (P_k^{-1} + \Phi_k \Phi_k^T) \mathbf{c}_k P_k \Phi_k W_{k+1} \\ & = W_{k+1}^T \Phi_k^T P_k \mathbf{c}_k^2 \Phi_k W_{k+1} \\ & \quad + W_{k+1}^T \Phi_k^T P_k \mathbf{c}_k \Phi_k \Phi_k^T \mathbf{c}_k P_k \Phi_k W_{k+1} \\ & = W_{k+1}^T \mathbf{b}_k^2 \Phi_k^T P_k \Phi_k W_{k+1} \\ & \quad + W_{k+1}^T (I_n + \Phi_k^T P_k \Phi_k) \mathbf{b}_k^2 \Phi_k^T P_k \Phi_k W_{k+1} \\ & \quad - W_{k+1}^T \mathbf{b}_k^2 \Phi_k^T P_k \Phi_k W_{k+1} \\ & = W_{k+1}^T \mathbf{b}_k \Phi_k^T P_k \Phi_k W_{k+1}. \quad (32) \end{aligned}$$

By (29), (31) and (32), we can get from (28) that

$$\begin{aligned} V_{k+1} & \leq V_k - \tilde{\Theta}_k^T \Phi_k \mathbf{b}_k \Phi_k^T \tilde{\Theta}_k - \tilde{\Theta}_k^T P_k^{-1} \Delta_{k+1} P_k^{-1} \tilde{\Theta}_k \\ & \quad - 2 \tilde{\Theta}_k^T P_k^{-1} \bar{P}_{k+1} \Phi_k W_{k+1} \\ & \quad + 2 \tilde{\Theta}_k^T P_k^{-1} \Delta_{k+1} \Phi_k W_{k+1} \\ & \quad + W_{k+1}^T \mathbf{b}_k \Phi_k^T P_k \Phi_k W_{k+1}. \quad (33) \end{aligned}$$

Summing from $k = 0$ to t yields

$$\begin{aligned} & V_{t+1} + \sum_{k=0}^t \tilde{\Theta}_k^T \Phi_k \mathbf{b}_k \Phi_k^T \tilde{\Theta}_k \\ & \quad + \sum_{k=0}^t \tilde{\Theta}_k^T P_k^{-1} \Delta_{k+1} P_k^{-1} \tilde{\Theta}_k \\ & \leq V_0 - 2 \sum_{k=0}^t \tilde{\Theta}_k^T P_k^{-1} \bar{P}_{k+1} \Phi_k W_{k+1} \\ & \quad - 2 \sum_{k=0}^t \tilde{\Theta}_k^T P_k^{-1} (-\Delta_{k+1}) \Phi_k W_{k+1} \\ & \quad + \sum_{k=0}^t W_{k+1}^T \mathbf{b}_k \Phi_k^T P_k \Phi_k W_{k+1}. \quad (34) \end{aligned}$$

Next, we estimate the last three terms on the RHS of (34) separately. By *Condition 3.1*, and $\tilde{\Theta}_k^T P_k^{-1} \bar{P}_{k+1} \Phi_k \in$

$\mathcal{F}_k, \tilde{\Theta}_k^T P_k^{-1} (-\Delta_{k+1}) \Phi_k \in \mathcal{F}_k$, we can use the martingale estimation theorem (*Theorem 2.8* in [41]) to get the following estimation for any $\delta > 0$,

$$\begin{aligned} & \sum_{k=0}^t \tilde{\Theta}_k^T P_k^{-1} \bar{P}_{k+1} \Phi_k W_{k+1} \\ &= O\left(\left\{\sum_{k=0}^t \|\tilde{\Theta}_k^T P_k^{-1} \bar{P}_{k+1} \Phi_k\|^2\right\}^{\frac{1}{2}+\delta}\right) \quad a.s., \end{aligned} \quad (35)$$

and

$$\begin{aligned} & \sum_{k=0}^t \tilde{\Theta}_k^T P_k^{-1} (-\Delta_{k+1}) \Phi_k W_{k+1} \\ &= O\left(\left\{\sum_{k=0}^t \|\tilde{\Theta}_k^T P_k^{-1} \Delta_{k+1} \Phi_k\|^2\right\}^{\frac{1}{2}+\delta}\right) \quad a.s. \end{aligned} \quad (36)$$

To further analyze (35) and (36), we note from (30) and the definitions of \bar{P}_{k+1} and b_k that

$$\begin{aligned} & P_k^{-1} \bar{P}_{k+1} \Phi_k \\ &= \Phi_k - c_k \Phi_k \Phi_k^T P_k \Phi_k \\ &= \Phi_k - c_k \Phi_k (I_n + \Phi_k^T P_k \Phi_k) + c_k \Phi_k \\ &= \Phi_k b_k. \end{aligned}$$

Hence, it is easy to see that

$$\begin{aligned} & \|\tilde{\Theta}_k^T P_k^{-1} \bar{P}_{k+1} \Phi_k\|^2 \\ &= \tilde{\Theta}_k^T P_k^{-1} \bar{P}_{k+1} \Phi_k \Phi_k^T \bar{P}_{k+1} P_k^{-1} \tilde{\Theta}_k \\ &= \tilde{\Theta}_k^T \Phi_k b_k^2 \Phi_k^T \tilde{\Theta}_k \\ &\leq \tilde{\Theta}_k^T \Phi_k b_k \Phi_k^T \tilde{\Theta}_k, \end{aligned} \quad (37)$$

where for the last inequality we have used the fact that $b_k \leq I_n$. By taking $0 < \delta < \frac{1}{2}$, we have from (35) and (37) that

$$\begin{aligned} & \sum_{k=0}^t \tilde{\Theta}_k^T P_k^{-1} \bar{P}_{k+1} \Phi_k W_{k+1} \\ &= O(1) + o\left(\left\{\sum_{k=0}^t \|\tilde{\Theta}_k^T P_k^{-1} \bar{P}_{k+1} \Phi_k\|^2\right\}\right) \\ &= O(1) + o\left(\left\{\sum_{k=0}^t \tilde{\Theta}_k^T \Phi_k b_k \Phi_k^T \tilde{\Theta}_k\right\}\right) \quad a.s. \end{aligned} \quad (38)$$

To further analyze (36), we now prove that

$$\Delta_{k+1} \Phi_k \Phi_k^T \Delta_{k+1} \leq \Delta_{k+1}. \quad (39)$$

For this, we need only to prove that

$$\Delta_{k+1}^{\frac{1}{2}} \Phi_k \Phi_k^T \Delta_{k+1}^{\frac{1}{2}} \leq I_{mn}.$$

Since $\Delta_{k+1} = \bar{P}_{k+1} - \mathcal{A} P_{k+1} \mathcal{A} \leq \bar{P}_{k+1}$, by *Lemma A.1* in Appendix A, we have

$$\begin{aligned} & \Delta_{k+1}^{\frac{1}{2}} \Phi_k \Phi_k^T \Delta_{k+1}^{\frac{1}{2}} \\ &\leq \lambda_{max}\{\Delta_{k+1}^{\frac{1}{2}} \Phi_k \Phi_k^T \Delta_{k+1}^{\frac{1}{2}}\} \cdot I_{mn} \\ &= \lambda_{max}\{\Phi_k^T \Delta_{k+1} \Phi_k\} \cdot I_{mn} \\ &\leq \lambda_{max}\{\Phi_k^T \bar{P}_{k+1} \Phi_k\} \cdot I_{mn} \end{aligned}$$

$$\begin{aligned} &= \lambda_{max}\{\Phi_k^T (P_k - c_k P_k \Phi_k \Phi_k^T P_k) \Phi_k\} \cdot I_{mn} \\ &= \lambda_{max}\{\Phi_k^T P_k \Phi_k - b_k (\Phi_k^T P_k \Phi_k)^2\} \cdot I_{mn} \\ &= \lambda_{max}\{b_k \Phi_k^T P_k \Phi_k\} \cdot I_{mn} < I_{mn}. \end{aligned}$$

Hence, we have (39), and so we have

$$\begin{aligned} & \|\tilde{\Theta}_k^T P_k^{-1} \Delta_{k+1} \Phi_k\|^2 \\ &= \tilde{\Theta}_k^T P_k^{-1} \Delta_{k+1} \Phi_k \Phi_k^T \Delta_{k+1} P_k^{-1} \tilde{\Theta}_k \\ &\leq \tilde{\Theta}_k^T P_k^{-1} \Delta_{k+1} P_k^{-1} \tilde{\Theta}_k. \end{aligned} \quad (40)$$

By taking $0 < \delta < \frac{1}{2}$, we know from (36) and (40) that

$$\begin{aligned} & \sum_{k=0}^t \tilde{\Theta}_k^T P_k^{-1} (-\Delta_{k+1}) \Phi_k W_{k+1} \\ &= O(1) + o\left(\left\{\sum_{k=0}^t \|\tilde{\Theta}_k^T P_k^{-1} \Delta_{k+1} \Phi_k\|^2\right\}\right) \\ &= O(1) + o\left(\left\{\sum_{k=0}^t \tilde{\Theta}_k^T P_k^{-1} \Delta_{k+1} P_k^{-1} \tilde{\Theta}_k\right\}\right) \quad a.s. \end{aligned} \quad (41)$$

We now proceed to estimate the last term in (34). Firstly, we know that

$$\begin{aligned} & W_{k+1}^T b_k \Phi_k^T P_k \Phi_k W_{k+1} \\ &\leq \|b_k \Phi_k^T P_k \Phi_k\| \cdot \|W_{k+1}\|^2 \\ &= \lambda_{max}\{b_k \Phi_k^T P_k \Phi_k\} \cdot \left\{\sum_{i=1}^n w_{k+1,i}^2\right\}. \end{aligned} \quad (42)$$

Following a similar proof idea in the traditional single sensor case ([38], see also [41]), from $\bar{P}_{k+1} = P_k - c_k P_k \Phi_k \Phi_k^T P_k$, we have $P_k^{-1} = \bar{P}_{k+1}^{-1} (I_{mn} - c_k P_k \Phi_k \Phi_k^T)$. By taking determinants on both sides of the above identity, and noticing $0 \leq b_k \Phi_k^T P_k \Phi_k \leq I_n$ and *Lemma A.1* in Appendix A, we have

$$\begin{aligned} |\mathbf{P}_k^{-1}| &= |\bar{\mathbf{P}}_{k+1}^{-1}| \cdot |I_{mn} - c_k P_k \Phi_k \Phi_k^T| \\ &= |\bar{\mathbf{P}}_{k+1}^{-1}| \cdot |I_n - b_k \Phi_k^T P_k \Phi_k| \\ &= |\bar{\mathbf{P}}_{k+1}^{-1}| \cdot \left\{\prod_{i=1}^n (1 - b_{k,i} \varphi_{k,i}^T P_{k,i} \varphi_{k,i})\right\} \\ &\leq |\bar{\mathbf{P}}_{k+1}^{-1}| \cdot (1 - \max_{i=1,\dots,n} \{b_{k,i} \varphi_{k,i}^T P_{k,i} \varphi_{k,i}\}) \\ &= |\bar{\mathbf{P}}_{k+1}^{-1}| \cdot (1 - \lambda_{max}\{b_k \Phi_k^T P_k \Phi_k\}). \end{aligned}$$

Moreover, we know from *Lemma 4.3* that

$$\begin{aligned} \lambda_{max}\{b_k \Phi_k^T P_k \Phi_k\} &\leq \frac{|\bar{\mathbf{P}}_{k+1}^{-1}| - |\mathbf{P}_k^{-1}|}{|\bar{\mathbf{P}}_{k+1}^{-1}|} \\ &= 1 - \frac{|\mathbf{P}_k^{-1}|}{|\bar{\mathbf{P}}_{k+1}^{-1}|} \\ &\leq 1 - \frac{|\mathbf{P}_k^{-1}|}{|\bar{\mathbf{P}}_{k+1}^{-1}|} \\ &\leq \frac{|\mathbf{P}_{k+1}^{-1}| - |\mathbf{P}_k^{-1}|}{|\bar{\mathbf{P}}_{k+1}^{-1}|}. \end{aligned}$$

Therefore

$$\begin{aligned} \sum_{k=0}^t \lambda_{\max}\{\mathbf{b}_k \Phi_k^T \mathbf{P}_k \Phi_k\} &\leq \sum_{k=0}^t \frac{|\mathbf{P}_{k+1}^{-1}| - |\mathbf{P}_k^{-1}|}{|\mathbf{P}_{k+1}^{-1}|} \\ &\leq \sum_{k=0}^t \int_{|\mathbf{P}_k^{-1}|}^{|\mathbf{P}_{k+1}^{-1}|} \frac{dx}{x} \\ &= \log(|\mathbf{P}_{t+1}^{-1}|) - \log(|\mathbf{P}_0^{-1}|). \end{aligned} \quad (43)$$

By the C_r -inequality and the Lyapunov inequality [60], it is easy to see that for any $\alpha \in (2, \min(\beta, 4))$,

$$\begin{aligned} &\sup_k \mathbb{E} \left[\left(\sum_{i=1}^n w_{k+1,i}^2 - \mathbb{E} \left[\sum_{i=1}^n w_{k+1,i}^2 \middle| \mathcal{F}_k \right] \right)^{\frac{\alpha}{2}} \middle| \mathcal{F}_k \right] \\ &\leq 2 \sup_k \mathbb{E} \left[\sum_{i=1}^n |w_{k+1,i}|^\alpha \middle| \mathcal{F}_k \right] < \infty, \quad a.s. \end{aligned}$$

Consequently, by using the martingale estimation theorem (*Theorem 2.8* in [41]), we have for any $\forall \eta > 0$,

$$\begin{aligned} &\sum_{k=0}^t \lambda_{\max}\{\mathbf{b}_k \Phi_k^T \mathbf{P}_k \Phi_k\} \\ &\cdot \left\{ \sum_{i=1}^n w_{k+1,i}^2 - \mathbb{E} \left[\sum_{i=1}^n w_{k+1,i}^2 \middle| \mathcal{F}_k \right] \right\} \\ &= O \left(S_t \left(\frac{\alpha}{2} \right) \left\{ \log \left(S_t \left(\frac{\alpha}{2} \right) + e \right) \right\}^{\frac{2}{\alpha} + \eta} \right), \quad a.s., \end{aligned} \quad (44)$$

where

$$S_t \left(\frac{\alpha}{2} \right) \triangleq \left[\sum_{k=0}^t (\lambda_{\max}\{\mathbf{b}_k \Phi_k^T \mathbf{P}_k \Phi_k\})^{\frac{\alpha}{2}} \right]^{\frac{2}{\alpha}}.$$

Since $\mathbf{b}_k \Phi_k^T \mathbf{P}_k \Phi_k \leq I_n$ and $\frac{\alpha}{2} > 1$, we have from (43) that

$$S_t \left(\frac{\alpha}{2} \right) = O(1) + O((\log |\mathbf{P}_{t+1}^{-1}|)^{\frac{2}{\alpha}}).$$

From this, we can get from (42)-(44) that

$$\begin{aligned} &\sum_{k=0}^t \mathbf{W}_{k+1}^T \mathbf{b}_k \Phi_k^T \mathbf{P}_k \Phi_k \mathbf{W}_{k+1} \\ &\leq \sum_{i=1}^n \sigma_i^2 \sum_{k=0}^t \lambda_{\max}\{\mathbf{b}_k \Phi_k^T \mathbf{P}_k \Phi_k\} + o(\log |\mathbf{P}_{t+1}^{-1}|) + O(1) \\ &\leq \sigma_w \log |\mathbf{P}_{t+1}^{-1}| + o(\log |\mathbf{P}_{t+1}^{-1}|) + O(1). \end{aligned}$$

Finally, substituting this together with (38) and (41) into (34), we know that the desired result (27) is true. This completes the proof. \blacksquare

Proof of Theorem 3.1.

Proof: By the definitions of $\bar{P}_{t,i}^{-1}$ and $P_{t,i}^{-1}$, it is easy to know that for any $t \geq 0$,

$$\begin{aligned} P_{t+1,i}^{-1} &= \sum_{j=1}^n a_{ji} \bar{P}_{t+1,j}^{-1} \\ &= \sum_{j=1}^n a_{ji} (P_{t,j}^{-1} + \varphi_{t,j} \varphi_{t,j}^T). \end{aligned} \quad (45)$$

Consequently, we have

$$\begin{aligned} &\max_{1 \leq i \leq n} \lambda_{\max}\{P_{t+1,i}^{-1}\} \\ &\leq \max_{1 \leq i \leq n} \sum_{j=1}^n a_{ji} \left(\lambda_{\max}\{P_{t,j}^{-1}\} + \lambda_{\max}\{\varphi_{t,j} \varphi_{t,j}^T\} \right) \\ &\leq \max_{1 \leq i \leq n} \lambda_{\max}\{P_{t,i}^{-1}\} \sum_{j=1}^n a_{ji} + \sum_{j=1}^n \lambda_{\max}\{\varphi_{t,j} \varphi_{t,j}^T\} \\ &= \max_{1 \leq i \leq n} \lambda_{\max}\{P_{t,i}^{-1}\} + \sum_{j=1}^n \|\varphi_{t,j}\|^2 \\ &\leq \dots \\ &\leq \max_{1 \leq i \leq n} \lambda_{\max}\{P_{0,i}^{-1}\} + \sum_{j=1}^n \sum_{k=0}^t \|\varphi_{k,j}\|^2 \\ &= \lambda_{\max}\{\mathbf{P}_0^{-1}\} + \sum_{j=1}^n \sum_{k=0}^t \|\varphi_{k,j}\|^2. \end{aligned} \quad (46)$$

From (46) and the connection between determinant and eigenvalues of a matrix, it is easy to conclude that

$$\begin{aligned} \log(|\mathbf{P}_{t+1}^{-1}|) &\leq mn \log \left(\max_{1 \leq i \leq n} \lambda_{\max}\{P_{t+1,i}^{-1}\} \right) \\ &\leq mn \log(r_t). \end{aligned} \quad (47)$$

Consequently, *Theorem 3.1* follows from this and *Lemma 4.4* immediately. \blacksquare

B. Proof of Theorem 3.2

By the definition of \mathbf{b}_k in (16), we know that

$$\Phi_k \Phi_k^T = \Phi_k \mathbf{b}_k \Phi_k^T + \Phi_k (\mathbf{b}_k \Phi_k^T \mathbf{P}_k \Phi_k) \Phi_k^T.$$

Then by noticing that \mathbf{b}_k , Φ_k and \mathbf{P}_k are (block) diagonal matrices, and $\Phi_k^T \mathbf{P}_k \Phi_k = O(1)$, *a.s.*, we know that

$$\begin{aligned} &\sum_{i=1}^n \sum_{k=0}^t R_{k,i} \\ &= \sum_{i=1}^n \sum_{k=0}^t (\varphi_{k,i}^T \tilde{\theta}_{k,i})^2 \\ &= \sum_{k=0}^t \tilde{\Theta}_k^T \Phi_k \Phi_k^T \tilde{\Theta}_k \\ &= \sum_{k=0}^t \tilde{\Theta}_k^T \Phi_k \mathbf{b}_k \Phi_k^T \tilde{\Theta}_k + \sum_{k=0}^t \tilde{\Theta}_k^T \Phi_k (\mathbf{b}_k \Phi_k^T \mathbf{P}_k \Phi_k) \Phi_k^T \tilde{\Theta}_k \\ &= \sum_{k=0}^t \tilde{\Theta}_k^T \Phi_k \mathbf{b}_k \Phi_k^T \tilde{\Theta}_k + \sum_{k=0}^t \tilde{\Theta}_k^T \Phi_k \mathbf{b}_k^{\frac{1}{2}} (\Phi_k^T \mathbf{P}_k \Phi_k) \mathbf{b}_k^{\frac{1}{2}} \Phi_k^T \tilde{\Theta}_k \\ &= O \left(\sum_{k=0}^t \tilde{\Theta}_k^T \Phi_k \mathbf{b}_k \Phi_k^T \tilde{\Theta}_k \right). \end{aligned} \quad (48)$$

Substituting this into *Theorem 3.1* 1), we conclude that (19) holds.

C. Proof of Theorem 3.3

For ease of representation, let $a_{ij}^{(s)}$ be the (i, j) -th entry of the matrix \mathcal{A}^s , $s \geq 1$. Note that $a_{ij}^{(1)} = a_{ij}$. By *Condition 3.2*

and *Remark 3.1*, we know that $a_{ji}^{(D_G)} \geq a_{min} > 0$, where $a_{min} = \min_{i,j \in \mathcal{V}} a_{ij}^{(D_G)} > 0$, and D_G is the diameter of the graph \mathcal{G} . Consequently, it is not difficult to see that for any $k > D_G$, $a_{ji}^{(k)} \geq a_{min}$ holds.

By (16), it is easy to see that for any $t \geq 0$,

$$\begin{aligned} & \text{vec}\{\mathbf{P}_{t+1}^{-1}\} \\ &= \mathcal{A} \text{vec}\{\bar{\mathbf{P}}_{t+1}^{-1}\} \\ &= \mathcal{A} \text{vec}\{\mathbf{P}_t^{-1}\} + \mathcal{A} \text{vec}\{\Phi_t \Phi_t^T\} \\ &= \dots \\ &= \mathcal{A}^{t+1} \text{vec}\{\mathbf{P}_0^{-1}\} + \sum_{k=0}^t \mathcal{A}^{t-k+1} \text{vec}\{\Phi_k \Phi_k^T\}, \end{aligned} \quad (49)$$

which implies that for any $t \geq D_G$,

$$\begin{aligned} P_{t+1,i}^{-1} &= \sum_{j=1}^n a_{ji}^{(t+1)} P_{0,j}^{-1} + \sum_{j=1}^n \sum_{k=0}^t a_{ji}^{(t-k+1)} \varphi_{k,j} \varphi_{k,j}^T \\ &\geq \sum_{j=1}^n a_{ji}^{(t+1)} P_{0,j}^{-1} + \sum_{j=1}^n \sum_{k=0}^{t-D_G+1} a_{ji}^{(t-k+1)} \varphi_{k,j} \varphi_{k,j}^T \\ &\geq a_{min} \sum_{j=1}^n P_{0,j}^{-1} + a_{min} \sum_{j=1}^n \sum_{k=0}^{t-D_G+1} \varphi_{k,j} \varphi_{k,j}^T, \end{aligned} \quad (50)$$

holds. From this, we conclude that

$$\lambda_{min}\{\mathbf{P}_{t+1}^{-1}\} \geq a_{min} \lambda_{min} \left\{ \sum_{j=1}^n P_{0,j}^{-1} + \sum_{j=1}^n \sum_{k=0}^{t-D_G+1} \varphi_{k,j} \varphi_{k,j}^T \right\}.$$

Note also that

$$\|\tilde{\Theta}_{t+1}\|^2 \leq \tilde{\Theta}_{t+1}^T \left[\frac{\mathbf{P}_{t+1}^{-1}}{\lambda_{min}\{\mathbf{P}_{t+1}^{-1}\}} \right] \tilde{\Theta}_{t+1}. \quad (51)$$

Hence, by (51), and 2) in *Theorem 3.1*, we know that *Theorem 3.3* holds.

V. CONCLUDING REMARKS

In this paper, we have established a convergence theory for a basic class of distributed LS algorithms, under quite general conditions on the measured information or data used in the estimation. The accumulated regret of adaptive predictors has been shown to have a celebrated logarithm increase without any excitation condition imposed on the system data, and the convergence rate of the distributed LS estimates has also been established under a cooperative excitation condition, which can be regarded as an extension of the weakest possible excitation condition known for the convergence of the classical LS. Neither independence and stationarity, nor Gaussian property, are required in our results, which makes it possible for our theory to be applicable to feedback control systems, and to lay a foundation for further investigation on related problems concerning the combination of learning, communication and control. Moreover, the cooperative excitation condition introduced and used in the paper indicates that the distributed LS can fulfill the estimation task cooperatively, even if any individual sensor cannot due to lack of necessary excitation. Of course, there are still a number of interesting problems for

further research, for examples, to consider other distributed estimation algorithms including ones based on forgetting factor LS or Kalman filter for tracking unknown time-varying signals (e.g. [32]), to investigate the case where both of the measurements and regressors contain noises (e.g. [61]), and to combine distributed learning with distributed control problems, etc.

APPENDIX A SOME BASIC LEMMAS

Lemma A.1. [60] Let $A \in \mathbb{R}^{d \times s}$ and $B \in \mathbb{R}^{s \times d}$ be two matrices. Then the nonzero eigenvalues of the matrices AB and BA are the same, and $|I_d + AB| = |I_s + BA|$ holds. Moreover, if $d = s$, then $|AB| = |A| \cdot |B| = |BA|$, $\text{Tr}(A) = \text{Tr}(A^T)$, $\text{Tr}(AB) = \text{Tr}(BA)$. Furthermore, if A and B are positive definite matrices with $A \geq B$, then $A^{-1} \leq B^{-1}$.

Lemma A.2. (Ky Fan Convex Theorem) [48] For any non-negative definite matrices $A_i \in \mathbb{R}^{m \times m}$ ($i = 1, \dots, n$), and any constants $0 \leq \lambda_i \leq 1$ ($i = 1, \dots, n$) satisfying $\sum_{i=1}^n \lambda_i = 1$, the following inequality holds:

$$|\lambda_1 A_1 + \lambda_2 A_2 + \dots + \lambda_n A_n| \geq |A_1|^{\lambda_1} |A_2|^{\lambda_2} \dots |A_n|^{\lambda_n}.$$

We remark that this lemma is exactly Lemma 1 in [48] for $n = 2$. For $n > 2$, this lemma can be proved easily by induction.

Lemma A.3. [60] For any matrices A, B, C and D with suitable dimensions,

$$(A + BDC)^{-1} = A^{-1} - A^{-1}B(D^{-1} + CA^{-1}B)^{-1}CA^{-1},$$

holds, provided that the relevant matrices are invertible.

Lemma A.4. [60] For any scalar sequence $a_j \geq 0$, ($j = 1, \dots, m$), the following C_r -inequality holds:

$$\left(\sum_{j=1}^m a_j \right)^r \leq \begin{cases} m^{r-1} \sum_{j=1}^m a_j^r, & r \geq 1, \\ \sum_{j=1}^m a_j^r, & 0 \leq r \leq 1. \end{cases}$$

APPENDIX B PROOF OF REMARK 3.3

Similar to the proof of *Lemma 4.4*, here we consider the following Lyapunov function:

$$\bar{V}_k = \mathbb{E}[\tilde{\Theta}_k^T \mathbf{P}_k^{-1} \tilde{\Theta}_k].$$

Since $\Delta_{k+1} = \bar{\mathbf{P}}_{k+1} - \mathcal{A} \mathbf{P}_{k+1} \mathcal{A} \geq 0$ and $\{\omega_{k,i}, \mathcal{F}_k\}$ is a martingale difference sequence, it is not difficult to convince oneself that one can take mathematical expectations on both sides of (33) to arrive at the following relationship:

$$\begin{aligned} \bar{V}_{k+1} &\leq \bar{V}_k - \mathbb{E}[\tilde{\Theta}_k^T \Phi_k \mathbf{b}_k \Phi_k^T \tilde{\Theta}_k] - \mathbb{E}[\tilde{\Theta}_k^T \mathbf{P}_k^{-1} \Delta_{k+1} \mathbf{P}_k^{-1} \tilde{\Theta}_k] \\ &\quad - 2\mathbb{E}[\tilde{\Theta}_k^T \mathbf{P}_k^{-1} \bar{\mathbf{P}}_{k+1} \Phi_k \mathbf{W}_{k+1}] \\ &\quad + 2\mathbb{E}[\tilde{\Theta}_k^T \mathbf{P}_k^{-1} \Delta_{k+1} \Phi_k \mathbf{W}_{k+1}] \\ &\quad + \mathbb{E}[\mathbf{W}_{k+1}^T \mathbf{b}_k \Phi_k^T \mathbf{P}_k \Phi_k \mathbf{W}_{k+1}] \\ &\leq \bar{V}_k - \mathbb{E}[\tilde{\Theta}_k^T \Phi_k \mathbf{b}_k \Phi_k^T \tilde{\Theta}_k] \\ &\quad - 2\mathbb{E}[\tilde{\Theta}_k^T \mathbf{P}_k^{-1} \mathcal{A} \mathbf{P}_{k+1} \mathcal{A} \Phi_k \mathbf{W}_{k+1}] \\ &\quad + \mathbb{E}[\mathbf{W}_{k+1}^T \mathbf{b}_k \Phi_k^T \mathbf{P}_k \Phi_k \mathbf{W}_{k+1}] \end{aligned}$$

$$\begin{aligned}
&= \bar{V}_k - \mathbb{E}[\tilde{\Theta}_k^T \Phi_k \mathbf{b}_k \Phi_k^T \tilde{\Theta}_k] \\
&\quad - 2\mathbb{E}[\mathbb{E}[\tilde{\Theta}_k^T \mathbf{P}_k^{-1} \mathcal{A} \mathbf{P}_{k+1} \mathcal{A} \Phi_k \mathbf{W}_{k+1} | \mathcal{F}_k]] \\
&\quad + \mathbb{E}[\mathbf{W}_{k+1}^T \mathbf{b}_k \Phi_k^T \mathbf{P}_k \Phi_k \mathbf{W}_{k+1}] \\
&= \bar{V}_k - \mathbb{E}[\tilde{\Theta}_k^T \Phi_k \mathbf{b}_k \Phi_k^T \tilde{\Theta}_k] \\
&\quad + \mathbb{E}[\mathbf{W}_{k+1}^T \mathbf{b}_k \Phi_k^T \mathbf{P}_k \Phi_k \mathbf{W}_{k+1}].
\end{aligned}$$

Similar to the proof of *Lemma 4.4* and *Theorem 3.1*, summing from $k = 0$ to t yields

$$\begin{aligned}
&\bar{V}_{t+1} + \sum_{k=0}^t \mathbb{E}[\tilde{\Theta}_k^T \Phi_k \mathbf{b}_k \Phi_k^T \tilde{\Theta}_k] \\
&\leq \bar{V}_0 + \sum_{k=0}^t \mathbb{E}[\mathbf{W}_{k+1}^T \mathbf{b}_k \Phi_k^T \mathbf{P}_k \Phi_k \mathbf{W}_{k+1}] \\
&\leq \bar{V}_0 + \mathbb{E}[\sigma_w \log(|\mathbf{P}_{t+1}^{-1}|)] - \mathbb{E}[\sigma_w \log(|\mathbf{P}_0^{-1}|)] \\
&\leq \bar{V}_0 + mn\bar{\sigma} \mathbb{E}[\log(r_t)] - \bar{\sigma} \mathbb{E}[\log(|\mathbf{P}_0^{-1}|)] \\
&\leq \bar{V}_0 + mn\bar{\sigma} \log(\mathbb{E}[r_t]) - \bar{\sigma} \mathbb{E}[\log(|\mathbf{P}_0^{-1}|)], \quad (52)
\end{aligned}$$

where for the last inequality we have used the fact that $\log(\cdot)$ is a concave function.

Since there exists a deterministic constant $c > 0$ such that $\|\Phi_t^T \mathbf{P}_t \Phi_t\| \leq c$, the following result holds by (48) and (52):

$$\begin{aligned}
&\sum_{i=1}^n \sum_{k=0}^t \mathbb{E}[R_{k,i}] \\
&= \sum_{k=0}^t \mathbb{E}[\tilde{\Theta}_k^T \Phi_k \Phi_k^T \tilde{\Theta}_k] \\
&\leq (1+c) \sum_{k=0}^t \mathbb{E}[\tilde{\Theta}_k^T \Phi_k \mathbf{b}_k \Phi_k^T \tilde{\Theta}_k] \\
&\leq (1+c) \left\{ mn\bar{\sigma} \log(\mathbb{E}[r_t]) + \mathbb{E}[\tilde{\Theta}_0^T \mathbf{P}_0^{-1} \tilde{\Theta}_0] \right. \\
&\quad \left. - \bar{\sigma} \mathbb{E}[\log(|\mathbf{P}_0^{-1}|)] \right\}.
\end{aligned}$$

This completes the proof.

REFERENCES

- [1] M. Taj and A. Cavallaro, "Distributed and decentralized multicamera tracking," *IEEE Signal Process. Mag.*, vol. 28, no. 3, pp. 46–58, May 2011.
- [2] A. H. Sayed, S. Y. Tu, J. Chen, X. Zhao, and Z. J. Towfic, "Diffusion strategies for adaptation and learning over networks: an examination of distributed strategies and network behavior," *IEEE Signal Process. Mag.*, vol. 30, no. 3, pp. 155–171, May 2013.
- [3] A. H. Sayed, "Adaptive networks," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 460–497, April 2014.
- [4] A. Khalili, A. Rastegarnia and S. Sanei, "Performance analysis of incremental LMS over flat fading channels," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 3, pp. 489–498, Sept. 2017.
- [5] A. H. Sayed and C. G. Lopes, "Distributed recursive least-squares strategies over adaptive networks," in *Proc. 40th Asilomar Conf. Signals, Syst. Comput.*, Pacific, Grove, CA, Oct. 2006, pp. 233–237.
- [6] J. P-Chaves, N. Bogdanovic, and K. Berberidis, "Distributed incremental-based RLS for node-specific parameter estimation," *IEEE Transactions on Signal Processing*, vol. 62, no. 20, pp. 5382–5397, Oct. 2014.
- [7] S. Y. Xie and L. Guo, "A necessary and sufficient condition for stability of LMS-based consensus adaptive filters," *Automatica*, vol. 93, pp. 12–19, July 2018.
- [8] S. Y. Xie and L. Guo, "Analysis of normalized least mean squares-based consensus adaptive filters under a general information condition," *SIAM J. on Control and Optimization*, vol. 56, no. 5, pp. 3404–3431, Sept. 2018.
- [9] G. Mateos and G. B. Giannakis, "Distributed recursive least-squares: stability and performance analysis," *IEEE Transactions on Signal Processing*, vol. 60, no. 7, pp. 3740–3754, July 2012.
- [10] G. Mateos, I. D. Schizas, and G. B. Giannakis, "Distributed recursive least-squares for consensus-based in-network adaptive estimation," *IEEE Transactions on Signal Processing*, vol. 57, no. 11, pp. 4583–4588, Nov. 2009.
- [11] C. Gratton, N. K. D. Venkateswoda, R. Arablouei and S. Werner, "Consensus-based distributed total least-squares estimation using parametric semidefinite programming," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom, 2019, pp. 5227–5231.
- [12] R. Carli, A. Chiuso, L. Schenato, and S. Zampieri, "Distributed Kalman filtering based on consensus strategies," *IEEE Journal on Selected Areas in Communications*, vol. 26, no. 4, pp. 622–633, May 2008.
- [13] G. Battistelli and L. Chisci, "Kullback-Leibler average, consensus on probability densities, and distributed state estimation with guaranteed stability," *Automatica*, vol. 50, no. 3, pp. 707–718, March 2014.
- [14] G. Battistelli, L. Chisci, G. Mugnai, A. Farina, and A. Graziano, "Consensus-based linear and nonlinear filtering," *IEEE Transactions on Automatic Control*, vol. 60, no. 5, pp. 1410–1415, May 2015.
- [15] Q. Liu, Z. Wang, X. He, and D. H. Zhou, "On Kalman-consensus filtering with random link failures over sensor networks," *IEEE Transactions on Automatic Control*, vol. 63, no. 8, pp. 2701–2708, Aug. 2018.
- [16] S. Das and J. Moura, "Consensus+innovations distributed Kalman filter with optimized gains," *IEEE Transactions on Signal Processing*, vol. 65, no. 2, pp. 467–481, Jan. 2017.
- [17] S. Y. Xie and L. Guo, "Analysis of distributed adaptive filters based on diffusion strategies over sensor networks," *IEEE Transactions on Automatic Control*, vol. 63, no. 11, pp. 3643–3658, Nov. 2018.
- [18] M. J. Piggott and V. Solo, "Diffusion LMS with correlated regressors i: realization-wise stability," *IEEE Transactions on Signal Processing*, vol. 64, no. 21, pp. 5473–5484, Nov. 2016.
- [19] H. Nosrati, M. Shamsi, S. M. Taheri and M. H. Sedaaghi, "Adaptive networks under non-stationary conditions: formulation, performance analysis, and application," *IEEE Transactions on Signal Processing*, vol. 63, no. 16, pp. 4300–4314, Aug. 2015.
- [20] S. Y. Tu and A. H. Sayed, "Diffusion strategies outperform consensus strategies for distributed estimation over adaptive networks," *IEEE Transactions on Signal Processing*, vol. 60, no. 12, pp. 6217–6234, Dec. 2012.
- [21] V. Vahidpour, A. Rastegarnia, A. Khalili, W. M. Bazzi and S. Sanei, "Analysis of partial diffusion LMS for adaptive estimation over networks with noisy links," *IEEE Transactions on Network Science and Engineering*, vol. 5, no. 2, pp. 101–112, April-June 2018.
- [22] I. E. K. Harrane, R. Flamary and C. Richard, "On reducing the communication cost of the diffusion LMS algorithm," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 5, no. 1, pp. 100–112, March 2019.
- [23] L. Xiao, S. Boyd, and S. Lall, "A space-time diffusion scheme for peer-to-peer least-squares estimation," in *Proceedings of 5th International Conference on Information Processing in Sensor Networks (IPSN 2006)*, Nashville, TN, April 2006, pp. 168–176.
- [24] F. S. Cattivelli, C. G. Lopes, and A. H. Sayed, "Diffusion recursive least squares for distributed estimation over adaptive networks," *IEEE Transactions on Signal Processing*, vol. 56, no. 5, pp. 1865–1877, May 2008.
- [25] A. Bertrand, M. Moonen, and A. H. Sayed, "Diffusion bias-compensated RLS estimation over adaptive networks," *IEEE Transactions on Signal Processing*, vol. 59, no. 11, pp. 5212–5224, Nov. 2011.
- [26] R. Arablouei, K. Dogancay, S. Werner, and Y-F Huang, "Adaptive distributed estimation based on recursive least-squares and partial diffusion," *IEEE Transactions on Signal Processing*, vol. 62, no. 14, pp. 3510–3522, July 2014.
- [27] V. Vahidpour, A. Rastegarnia, A. Khalili and S. Sanei, "Analysis of partial diffusion recursive least squares adaptation over noisy links," *IET Signal Processing*, vol. 11, no. 6, pp. 749–757, August 2017.
- [28] A. Rastegarnia, "Reduced-communication diffusion RLS for distributed estimation over multi-agent networks," *IEEE Transactions on Circuits and Systems II: Express Briefs*, doi: 10.1109/TCSII.2019.2899194.
- [29] Y. Yu, H. Zhao, R. C. de Lamare, Y. Zakharov and L. Lu, "Robust distributed diffusion recursive least squares algorithms with side information for adaptive networks," *IEEE Transactions on Signal Processing*, vol. 67, no. 6, pp. 1566–1581, March 2019.
- [30] B. Widrow and S. Stearns, *Adaptive Signal Processing*, Prentice-Hall Englewood Cliffs NJ, 1985.

- [31] V. Solo and X. Kong, *Adaptive Signal Processing Algorithms*, Prentice Hall, 1995.
- [32] L. Guo, "Stability of recursive stochastic tracking algorithms," *SIAM J. on Control and Optimization*, vol. 32, no. 5, pp. 1195–1225, Sept. 1994.
- [33] L. Guo, "Convergence and logarithm laws of self-tuning regulators," *Automatica*, vol. 31, no. 3, pp. 435–450, March 1995.
- [34] K. J. Åström and B. Wittenmark, "On self tuning regulators," *Automatica*, vol. 9, no. 2, pp. 185–199, March 1973.
- [35] L. Ljung, "Consistency of the least-squares identification method," vol. 21, no. 5, pp. 779–781, Oct. 1976.
- [36] J. B. Moore, "On strong consistency of least squares identification algorithm," *Automatica*, vol. 14, no. 5, pp. 505–509, Sep. 1978.
- [37] H. F. Chen, "Strong consistency and convergence rate of least squares identification," *Sci. Sinica*, Ser. A 25, no. 7, pp. 771–784, 1982.
- [38] T. L. Lai and C. Z. Wei, "Least squares estimates in stochastic regression models with applications to identification and control dynamic systems," *Annals of Statistics*, vol. 10, no. 1, pp. 154–166, March 1982.
- [39] T. L. Lai and C. Z. Wei, "Extended least squares and their applications to adaptive control and prediction in linear systems," *IEEE Transactions on Automatic Control*, vol. 31, no. 10, pp. 898–906, Oct. 1986.
- [40] H. F. Chen and L. Guo, "Convergence rate of least-squares identification and adaptive control for stochastic systems," *int. J. Control*, vol. 44, no. 5, pp. 1459–1476, Nov. 1986.
- [41] H. F. Chen and L. Guo, *Identification and Stochastic Adaptive Control*, Birkhäuser, Boston, 1991.
- [42] L. Guo and H. F. Chen, "The Åström-Wittenmark self-tuning regulator revised and ELS-based adaptive tracker," *IEEE Transactions on Automatic Control*, vol. 36, no. 7, pp. 802–812, July 1991.
- [43] L. Ljung, "Analysis of recursive stochastic algorithms," *IEEE Transactions on Automatic Control*, vol. 22, no. 4, pp. 551–575, Aug. 1977.
- [44] G. C. Goodwin, P. J. Ramadge, and P. E. Caines, "Discrete time stochastic adaptive control," *SIAM J. on Control and Optimization*, vol. 19, no. 6, pp. 829–853, 1981.
- [45] P. R. Kumar, "Convergence of adaptive control schemes using least-squares parameter estimates," *IEEE Transactions on Automatic Control*, vol. 35, no. 4, pp. 416–424, April 1990.
- [46] S. Julier and J. Uhlmann, "A non-divergent estimation algorithm in the presence of unknown correlations," in *Proceedings of the 1997 American Control Conference*, Albuquerque, NM, USA, April 1997, pp. 2369–2373.
- [47] L. Chen, P. Arambel, and P. Mehra, "Estimation under unknown correlation: covariance intersection revisited," *IEEE Transactions on Automatic Control*, vol. 47, no. 11, pp. 1879–1882, May 2002.
- [48] Ky Fan, "On a theorem of Weyl concerning eigenvalues of linear transformations," *Proc. Nat. Acad. USA*, vol. 36, no. 1, pp. 31–35, Jan. 1950.
- [49] Y. S. Chow and H. Teicher, *Probability Theory*, New York: Springer, March 1978.
- [50] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM Review*, vol. 60, no. 2, pp. 223–311, May 2018.
- [51] F. Yan, S. Sundaram, S. V. N. Vishwanathan, and Y. Qi, "Distributed autonomous online learning: regrets and intrinsic privacy-preserving properties," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 11, pp. 2483–2493, Nov. 2013.
- [52] S. Hosseini, A. Chapman, and M. Mesbahi, "Online distributed convex optimization on dynamic networks," *IEEE Transactions on Automatic Control*, vol. 61, no. 11, pp. 3545–3550, Nov. 2016.
- [53] M. Akbari, B. Ghahesifard, and T. Linder, "Distributed online convex optimization on time-varying directed graphs," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 3, pp. 417–428, Sep. 2017.
- [54] S. Shahrampour and A. Jadbabaie, "Distributed online optimization in dynamic environments using mirror descent," *IEEE Transactions on Automatic Control*, vol. 63, no. 3, pp. 714–725, March 2018.
- [55] M. Zhong and C. G. Cassandras, "Asynchronous distributed optimization with event-driven communication," *IEEE Transactions on Automatic Control*, vol. 55, no. 12, pp. 2735–2750, Dec. 2010.
- [56] S. Y. Xie and L. Guo, "Analysis of compressed distributed adaptive filters," provisionally accept by *Automatica*, 2019.
- [57] C. Godsil and G. Royle, *Algebraic Graph Theory*, Springer-Verlag, 2014.
- [58] T. L. Lai, "Asymptotically efficient adaptive control in stochastic regression models," *Advances in Applied Mathematics*, vol. 7, no. 1, pp. 23–45, March 1986.
- [59] L. Ljung and L. Guo, "The role of model validation for assessing the size of the unmodeled dynamics," *IEEE Transactions on automatic control*, vol. 42, no. 9, pp. 1230–1239, Sep. 1997.
- [60] L. Guo, *Time-Varying Stochastic Systems—Stability, Estimation and Control*, Jilin Science and Technology Press, 1993.
- [61] W. X. Zheng, "On least-squares identification of stochastic linear systems with noisy input-output data," *International Journal of Adaptive Control and Signal Processing*, vol. 13, no. 3, pp. 131–143, 1999.